Evolutionary Trajectories of IDH Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis

Körber, Verena ; Yang, Jing ; Barah, Pankaj ; Wu, Yonghe ; Stichel, Damian ; Gu, Zuguang ; et al ; Weller, Michael

Abstract: We studied how intratumoral genetic heterogeneity shapes tumor growth and therapy response for isocitrate dehydrogenase (IDH)-wild-type glioblastoma, a rapidly regrowing tumor. We inferred the evolutionary trajectories of matched pairs of primary and relapsed tumors based on deep whole-genome-sequencing data. This analysis suggests both a distant origin of de novo glioblastoma, up to 7 years before diagnosis, and a common path of early tumorigenesis, with one or more of chromosome 7 gain, 9p loss, or 10 loss, at tumor initiation. TERT promoter mutations often occurred later as a prerequisite for rapid growth. In contrast to this common early path, relapsed tumors acquired no stereotypical pattern of mutations and typically regrew from oligoclonal origins, suggesting sparse selective pressure by therapeutic measures.

DOI: https://doi.org/10.1016/j.ccell.2019.02.007

Evolutionary trajectories of IDH-wildtype glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis

Verena Körber[1,2], Jing Yang[3,4], Pankaj Barah[3], Yonghe Wu[5,7], Damian Stichel[6], Zuguang Gu[3,7], Michael Nai Chung Fletcher[5], David Jones[8], Bettina Hentschel[9], Katrin Lamszus[10], Jörg Christian Tonn[11], Gabriele Schackert[12], Michael Sabel[13], Jörg Felsberg[14], Angela Zacher[14], Kerstin Kaulich[14], Daniel Hübschmann[3,15,16], Christel Herold-Mende[17], Andreas von Deimling[6], Michael Weller[18], Bernhard Radlwimmer[5], Matthias Schlesner[4], Guido Reifenberger[14]*, Thomas Höfer[1,2]*, Peter Lichter[5,7,19]*

[1]Division of Theoretical Systems Biology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

[2]Bioquant Center, Heidelberg University, Im Neuenheimer Feld 267, D-69120 Heidelberg, Germany

[3]Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

[4]Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

[5]Division of Molecular Genetics, German Cancer Research Consortium (DKTK), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

[6]Clinical Cooperation Unit Neuropathology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

[7]Heidelberg Center for Personalized Oncology, DKFZ-HIPO, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

[8]Pediatric Glioma Research, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

[9]Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Härtelstr. 16-18, D-04107 Leipzig, Germany

[10]Department of Neurosurgery, University Medical Center Hamburg-Eppendorf, Neues Klinikum O10, Martinistr. 52, D-20246 Hamburg, Germany

[11]Department of Neurosurgery, Ludwig Maximilians University Munich and German Cancer Consortium (DKTK), partner site Munich, Marchioninistraße 15, D-81377 Munich, Germany

[12]Department of Neurosurgery, Technical University Dresden, Fetscherstr. 74, D-01307 Dresden, Germany

[13]Department of Neurosurgery, Heinrich Heine University Düsseldorf, Moorenstr. 5, D-40255 Düsseldorf, Germany

[14]Institute of Neuropathology, Heinrich Heine University Düsseldorf, and German Cancer Consortium (DKTK), partner site Essen/Düsseldorf, Moorenstr. 5, D-40255 Düsseldorf, Germany
[15]Department of Pediatric Immunology, Hematology and Oncology, Heidelberg University Hospital, Im Neuenheimer Feld 410, D-69120 Heidelberg, Germany
[16]Division of Stem Cells and Cancer, German Cancer Research Center (DKFZ) and Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM gGmbH), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany
[17]Department of Neurosurgery, Heidelberg University Hospital, Im Neuenheimer Feld 400, D-69120 Heidelberg, Germany.
[18]Department of Neurology, University Hospital Zurich, Frauenklinikstr. 26, CH-8091 Zurich, Switzerland
[19]Lead contact
*Corresponding authors: Guido.Reifenberger@med.uni-duesseldorf.de, t.hoefer@dkfz.de, peter.lichter@dkfz.de

Summary

We studied how intratumoral genetic heterogeneity shapes tumor growth and therapy response for isocitrate dehydrogenase (IDH)-wildtype glioblastoma, a rapidly regrowing tumor. To this end, we inferred the evolutionary trajectories of matched pairs of primary/relapsed tumors based on deep whole-genome-sequencing. This analysis suggests both a distant origin of *de novo* glioblastoma, up to seven years before diagnosis, and a common path of early tumorigenesis, with one or more of three events (chromosome 7 gain, 9p loss or 10 loss) at tumor initiation. *TERT* promoter mutations often occurred later as a prerequisite for rapid growth. In contrast to this common early path, relapsed tumors acquired no stereotypical pattern of mutations and typically regrew from oligoclonal origins, suggesting sparse selective pressure by therapeutic measures.

Significance

Many tumors consist of genetically diverse subclones, which are thought to reflect cancer evolution. Indeed, we uncover large genetic diversity in tissue samples of aggressive brain tumors, IDH-wildtype glioblastomas, obtained at primary and recurrent surgical resection. Surprisingly, however, all these diverse tumors (which also belong to the known four subgroups) map onto a common path of early tumorigenesis where characteristic driver mutations are acquired by losses or gains of (parts of) chromosomes. Mutation rates suggest that these tumor-initiating events occur several years before diagnosis. Other common drivers, including *TERT* promoter mutations, often follow and may allow the tumor to grow to detectable size. Further genetic diversification appears to contribute little to the regrowth of the tumors after therapy.

Introduction

Tumors are typically heterogeneous mixtures of genetic subclones. Analyses of subclonal tumor evolution have focused on oncogenic drivers, which may inform the search for more effective therapeutic approaches (Greaves and Maley, 2012). However, recent studies across several cancer entities (Williams et al., 2016) or across many locations within one tumor (Ling et al., 2015) have questioned this idea by arguing that intra-tumoral genetic heterogeneity is often due to neutral (passenger) mutations. As tumor growth as a whole is caused by driver mutations, these findings concern the fundamental, but poorly understood, question of the relation between tumor growth and genetic evolution.

Here we address this question for the genetic evolution of glioblastoma. Glioblastomas are the most frequent malignant primary brain cancers in adults that rapidly recur after initial therapy and thus rank among the human malignancies with the highest mortality (with a 5-year survival rate of only 6%; Ostrom et al., 2016). Thus, it is of particular clinical interest whether standard chemoradiotherapy exerts selective pressure on the evolution of the relapse tumor. With advances in whole-genome sequencing, genetically distinct subclones within glioblastomas have been identified. Several studies have suggested that branched evolution of the tumor contributes to treatment failure (Francis et al., 2014; Hu et al., 2017; Meyer et al., 2013; Patel et al., 2014; Snuderl et al., 2011; Sottoriva et al., 2013; Yates et al., 2015). In this view, therapeutic intervention, typically consisting of neurosurgical tumor resection followed by chemoradiotherapy with temozolomide (TMZ) (Weller et al., 2017) may select resistant subclones from residual tumor cells which quickly regrow (Gerlinger and Swanton, 2010; Nathanson et al., 2014; Qazi et al., 2017). Moreover, exome sequencing of pairs of isocitrate dehydrogenase (IDH)-mutant diffuse gliomas of World Health Organisation (WHO) grade II that recurred as WHO grade III or IV gliomas suggests that therapy may drive tumor evolution to more malignant stages, with 60% of patients receiving TMZ reported to show a strong increase in tumor mutation rate (Johnson et al., 2014). To investigate whether these scenarios of progressive tumor evolution driven by cytotoxic therapy also apply to IDH-wildtype tumors that were already glioblastomas at initial diagnosis (*de novo* glioblastomas), we

have utilized the distribution of clonal and subclonal mutations to infer both the genetic evolution and the growth history of *de novo* glioblastomas prior to and following standard therapy.

Results

Mutation patterns in primary and recurrent glioblastomas are largely similar

To assess the genetic evolution of glioblastomas, we analyzed the genomes, transcriptomes and methylomes in paired primary (untreated) and recurrent (following initial treatment) tumor tissue samples from patients with *de novo* glioblastomas, isocitrate dehydrogenase wildtype (IDH^{WT}), World Health Organization (WHO) grade IV. In total, paired tissue samples from 50 patients were included, the majority of whom had received radiotherapy combined with concomitant and maintenance temozolomide (TMZ) chemotherapy after neurosurgical resection (Figure 1A and Table S1). We chose tumor pairs from 21 patients as a discovery set and sequenced whole genomes of both samples (149x average coverage) and a matched blood control (78x average coverage), as well as the transcriptomes of primary and recurrent tumors (Figure 1B). The median tumor cell contents were 80% and 78% for primary and recurrent samples, respectively (Figure S1A). Transcriptome and methylome analyses indicated that the vast majority of non-tumor cells were infiltrating immune cells, particularly monocytes (Figure S1B,C) (cf. Wang et al., 2017). In addition, we sequenced a panel of 50 glioma-associated genes in primary and recurrent tumor pairs from 43 IDH^{WT} glioblastoma patients, including 14 of the 21 patients included in the whole genome sequencing (WGS) discovery set (Table S1). Assessing genome-wide DNA methylation patterns in both sets with Illumina 450k/EPIC bead chip arrays, we found the previously described major glioblastoma subgroups by unsupervised clustering (Capper et al., 2018; Sturm et al., 2012) (Figure 1B). We noted that samples classified as mesenchymal subtype had comparatively low tumor content, suggesting that this particular classification may be influenced by a high fraction of non-tumor cells. When excluding subtype switches involving the mesenchymal subtype, we found that

90% of primary and recurrent tumor pairs had a stable methylation subtype, while the remainder switched from the RTKII to the RTKI group (Figure 1B, Figure S1D).

Whole genome sequencing yielded a median mutational burden of 12,800 somatic single nucleotide variants (SNVs) and small insertions/deletions (indels) per tumor, and most samples had a mutation count in this order of magnitude (Figure 1C). One primary and four recurrent tumors had vastly more mutations ($>10^5$), which included somatic mutations in DNA mismatch repair genes (*MLH1* or *MSH6*), consistent with a hypermutation genotype. Apart from these hypermutated cases, primary and recurrent tumor samples had comparable numbers of mutations (p value = 0.77, two-sided Welch two sample t-test). On average, about as many mutations were shared between primary and recurrent samples as were private to either sample (Figures 1D and 1E). The shared mutations were dominated by the clock-like mutational signature 1 (Figure 1F; classification according to COSMIC; Alexandrov et al., 2013; Forbes et al., 2016). The pattern of private mutational signatures in both primary and non-hypermutated recurrent tumors were shifted away from signature 1 and included signatures due to defective double-strand break repair by homologous recombination (signature 3), defective mismatch repair (signature 15 and signature 26) and one of unknown etiology (signature 5) (Figure 1G). The hypermutated tumors were all dominated by mutational signature 11 (Figure 1H), described as being linked to TMZ treatment (Alexandrov et al., 2013). The similarity of mutation counts and signatures in paired primary and recurrent tumors suggests that the major part of the genetic evolution occurred prior to the diagnosis of the primary tumor. Only the comparatively small subgroup of hypermutated tumors showed a global mutational signature of cytotoxic treatment.

Paucity of common mutations in recurrent tumors

We identified recurrent driver mutations in SNVs, indels, copy number variations (CNVs) and structural variants (SVs) [several of which are druggable (Gröbner et al., 2018), Figure S2A, Methods]. Driver mutations detectable in tumors of at least three patients were found in the coding regions of 28 genes, in 13 non-coding RNAs and the *TERT* promoter region (Figure 2A). All

21 tumor pairs harbored at least one of three chromosomal gains or losses: (i) loss of chromosome 10 or the 10q arm, including the *PTEN* locus; (ii) loss of chromosome arm 9p or focal deletion of the *CDKN2A/B* locus; (iii) complete or partial gain of chromosome 7, including the *EGFR* locus (Figure 2A, B). The majority of 9p losses and focal deletions of *CDKN2A/B* were homozygous. Chromosome 10 or 10q loss was always hemizygous but in 75% of the tumor samples accompanied by a mutation in the remaining *PTEN* allele. Chromosome 7 gains were frequently (58%) accompanied by focal amplifications of the *EGFR* locus (which may be extrachromosomal in the form of double minutes; deCarvalho et al., 2018). When present, focal *EGFR* amplification occurred in all but one case already in the primary tumor and was lost in the recurrent samples in 27% of the cases. Moreover, almost half of the *EGFR* amplifications (44%) co-occurred with the active variant EGFRvIII (generated by deletion of exons 2-7; Brennan et al., 2013), and in three samples *EGFR* had an interchromosomal translocation into the vicinity of a superenhancer (Figure 2C). In several cases chromosomal gains increased (Figure 2D) or decreased (Figure 2E) in complexity between primary and recurrent tumors.

TERT promoter mutations were found in all but one primary sample (in the latter case, it was detected neither in the WGS nor the panel sequencing data). In contrast to the above chromosomal gains/losses and the *TERT* promoter mutations, the most frequent coding mutations (in *PTEN*, *EGFR* or *TP53*) were present in a subset of the tumors only. We observed comparable distributions of SNVs and small indels in these genes, as well as *TERT* promoter mutations, in the independent validation set (Figure S2B). Moreover, mutations in thirteen non-coding RNAs were recurrent (Figure 2A). The occurrence of widespread driver mutations in primary tumors that are maintained in the corresponding recurrent tumors (*PTEN*, *CDKN2A/B*, *EGFR*, *TERT*) contrasted with a paucity of common mutations newly detected in the recurrent tumor samples. Except for mutations in the mismatch repair gene *MSH6* and in the non-coding RNA *XIST*, which were enriched in the hypermutated recurrent samples, no recurrent mutations were preferentially found upon glioblastoma recurrence. Moreover, hypermutated tumors have a higher probability for recurrence of mutations by chance. There were, however, several cases of parallel evolution in both

the discovery and the validation set. We found different mutations in pairs of primary and recurrent samples for the following eight driver genes: *EGFR* (5 pairs), *ARID1A*, *AC005154.6*, *KCNQ1OT1*, *PTEN, PTPN11*, *TSIX*, *XIST* (one pair each).

Genetic evolution of glioblastoma shows a common pattern across DNA methylation subtypes

To understand how the mutation patterns in primary and recurrent glioblastomas arise, we inferred evolutionary histories for the individual sample pairs. We determined for each sample the allele frequencies of SNVs and indels (variant allele frequencies, VAFs) at non-amplified loci. Mutations present in both primary and recurrent tumor samples showed a wide range of VAFs (Figure 3A). Many shared mutations will have arisen comparatively early in tumor development and will therefore be clonal in both samples. Based on the read coverage of a given locus, we tested with a binomial model whether a measured VAF is compatible with sampling from a VAF of 50%, implying that the respective allele is clonal. Most VAFs were indeed compatible with clonality. While this test minimizes erroneous assignment of a mutation as subclonal, we found subclonality in both primary and recurrent samples for 12% of shared mutations (Figure 3A, red dots). An additional 7% of mutations were identified as clonal in the primary tumor and subclonal in the recurrent tumor (Figure 3A, orange dots). This outcome is not possible for a tumor with a single origin and hence was most likely due to undersampling of the primary tumor, effectively augmenting the count of subclonal mutations. Finally, a transition from subclonality in primary to clonality in the recurrent tumors is suggested for 11% of shared mutations (Figure 3A, blue dots). However, this outcome is unlikely to be the result of genetic evolution; it will rather encompass cases where the recurrent tumor has been undersampled (STAR Methods, Subclonality test) and few cases where surgical resection retained a particular subclone that then forms a clonal relapse tumor (see Figure 3H below). Taken together, the VAF analysis indicates extensive subclonality of shared mutations in pairs of primary and recurrent glioblastoma tissue samples.

Subclonality is informative on tumor evolution. To uncover this information, we mapped the observed intratumoral genetic heterogeneity onto the subclonal phylogeny of primary and recurrent tumors. We combined SNVs and CNVs to this end, developing a likelihood-based multinomial model that jointly infers genetically distinct subpopulations and their phylogenetic relationships from the sequencing read count distribution over all mutated loci (Figures 3B and S3A, B). We assessed the robustness of our algorithm on simulated data with characteristics of the measured data (150x coverage and tumor cell contents between 50% and 100%) and found reliable detection of even small clones (5% clone size detected with >80% sensitivity) and faithful inference of subclonal phylogenies (Figure S3C-L). We then applied subclonal inference to the measured data, treating normal tissue as an additional subclone, and found that the estimated tumor cell content closely agreed with purity estimates from ACEseq (*CC* = 0.95, Figure S3M). Thus our likelihood-based inference method performed well with simulated data and reproduced the independent estimate of tumor cell content in the actual data.

We determined the subclonal structure jointly for pairs of primary and recurrent tumor samples and typically resolved two to three subclones per sample (Figure 3C), with a clear hierarchy of clone sizes (Figure 3D). Allowing for up to five subclones per tumor sample did not alter the best-fit tree in all but two cases; in the latter two cases additional subclones were added within the already existing tree structure. Thus the resolution of up to three subclones per sample allowed reliable placement of driver mutations in tumor phylogenies.

In more than two thirds (15/21) of the sample pairs, the recurrent tumor originated from more than one clone of the primary tumor (oligoclonal origin of the relapse; Figure 3F,G and Figure S4A-N); this oligoclonal relapse scenario included all hypermutated cases. Figure 3E introduces the layout of the phylogenetic trees, and Figure 3F shows a typical example with oligoclonal origin of the relapse tumor. Note that, due to limited sampling, the evolution of a recurrent tumor subclone in the inferred phylogenetic trees does not start directly with the corresponding subclone in the primary tumor but, instead, both subclones have an unobserved common ancestor. In the example tree (Figure 3F), one allele of *PTEN* was lost (chromosome 10 loss) and the other one was

inactivated by a mutation, before subclonal branching became evident. Subclonal evolution in the primary tumor added further potential drivers, notably a partial gain of chromosome 7 (incl. *EGFR*) and mutations in the *TERT* promoter, *BRAF* (D594G) and *TSIX*. The recurrent tumor arose from two subclones of the primary tumor and showed further, subclone-specific, driver mutations. Notably, the subclones of the recurrent tumor acquired two different mutations in *LINC00343* in parallel. Figure 3G depicts an extreme case of oligoclonal relapse where no further recurrent drivers developed within 307 days between primary and recurrent resection of the tumor (median time span between initial diagnosis and relapse 284 days). In the minority of cases (6/21) the relapse tumor had a monoclonal origin (Figure 3H and Figure S4O-S). However, this might be partially due to taking experimental samples and, indeed, we found only one clear example of linear evolution of the recurrent tumor, with several new driver mutations being acquired over a comparatively large time span of ~22 months between primary and secondary resection (Figure 3H). In this case, the evolutionary time, measured by the number of mutations between the most recent common ancestor of the subclones in the primary tumor and the most recent common ancestor of the relapse subclones was indeed comparatively large (~5000 mutations). In stark contrast to this unique case, all other tumor pairs had identical, or very similar, evolutionary times of emergence of primary and relapse subclones (Figure 3I). Hence recurrent tumors mainly emerged from already established heterogeneity of the primary tumor and only moderately added to that. This common pattern of glioblastoma evolution was independent of DNA methylation subtype. Moreover, we specifically analyzed the methylation status of the *MGMT* promoter associated with response to TMZ therapy (Hegi et al., 2005) and found an apparently random distribution of methylated and unmethylated promoters between the two groups of oligoclonal and monoclonal relapse origins (i.e., 21% probability to choose the specific distribution observed at random). The common pattern of glioblastoma evolution allowed us to derive a consensus phylogeny over all analyzed tumors. To this end, we walked down the phylogenetic trees from the common stem and scaled branch thickness according to the number of branch occurrences over all trees, revealing a typical pattern of each primary subclone spawning a corresponding relapse subclone (Figure 3K).

The frequency of recurrent mutations declined rapidly from the common stem towards the branches, suggesting that therapeutic measures do not drive preferential resistance scenarios.

Copy number variations precede single-nucleotide variations as oncogenic drivers

Clonal mutations (present in all cells of a given primary and recurrent tumor sample) map to the common stem of the phylogenetic tree and, for each tumor, contain recurrent driver mutations. These drivers reached clonality due to their early origin in the evolution of the tumor. Further driver mutations were subclonal and hence originated more recently than the drivers in the stem. The majority of subclones in relapsed tumors (77%) showed additional driver mutations that were not found in the primary tumor samples, indicating ongoing genetic evolution.

We found strong overrepresentation of gain of chromosome 7, loss of chromosome arm 9p/focal loss of *CDKN2A/B* and loss of chromosome 10/10q in the common stem of the phylogenetic trees, with at least one of these chromosomal gains or losses being clonal in 20/21 tumor pairs (Figure 4A). The single tumor, where none of these chromosomal aberrations was clonal, harbored two recurrent mutations (as in Figure 2) at the clonal level, a frameshift insertion in *PTEN* and a clonal SNV in exon 7 of *TP53* (Pro>Ser; aa 278). The most prominent oncogenes on the altered chromosomes are *EGFR* (Chr. 7), *CDKN2A/B* (Chr. 9) and *PTEN* (Chr. 10), suggesting that mutations and/or copy number variations in these three genes define the minimal set of driver mutations from which at least one, and frequently two or all (81%), occur in the stem of each analyzed tumor. Hence these mutations might be early tumorigenic events.

In contrast to the CNVs, most recurrent SNVs were preferentially subclonal, with the exceptions of mutations in *TP53, PTEN, EGFR* and the *TERT* promoter (Figure 4B). Notably, most clonal *PTEN* mutations (7/10) co-occurred with a loss of the other *PTEN* allele. Mutations in the *TERT* promoter occurred in all but one of the 42 primary and recurrent samples, but, remarkably, were subclonal in 33% of them. Of note, the read coverage of the GC-rich *TERT* promoter was uniformly high (123x on average), allowing accurate VAF estimation; moreover, comparison of the VAF distribution of the *TERT* promoter mutations with a control set of heterozygous germline

mutations, including a germline SNP in the *TERT* promoter region, clearly showed significant subclonality in the former (Figure S5A-D). Nevertheless, in all but one case subclonal *TERT* mutations occurred in the primary tumor. In addition to canonical *TERT* promoter mutations reported previously, we also found exclusively 'non-canonical' promoter mutations in two tumors (Figure 4C). These tumors also had clonal mutations in *ATRX*, which inhibits alternative telomere lengthening (exonic frameshift generating a premature stop codon). Analyzing one such case in detail, we found that the *ATRX* mutation likely supported tumor growth, as the protein was lost (Figure S5F), whereas the non-canonical *TERT* promoter mutation did not increase transcript levels compared to non-tumor brain tissue (Figure S5G). These and further data (Figure S5H) are consistent with the view that *ATRX* mutations and canonical *TERT* promoter mutations are independent mechanisms to support survival of proliferating tumor cells.

Taken together, the placement of mutations on the phylogenetic trees indicates characteristic patterns of their occurrence: Copy number variations in chromosomes 7, 9 or 10 generally took place early in the evolution of the tumor while *TERT* promoter mutations were often acquired subsequently in primary tumors, remaining subclonal in one third of them. The absence of mutations that are characteristic for recurrent tumors may indicate that little directed selective pressure is exerted by chemotherapy and/or radiotherapy.

Large selective advantage of *TERT* promoter mutations

We reasoned that combining data on the number of somatic mutations accumulated during tumor evolution with estimates of the characteristic tumor size upon primary resection and the somatic mutation rate will yield insight into the evolution of IDH^{WT} glioblastoma. Tumor growth results from the balance of cell proliferation and death rates while the bulk of mutations is assumed to occur during cell proliferation (Figure 5A). A human brain tumor of 20 cubic centimeters or larger (Goldberg-Zimring et al., 2005) corresponds to at least 10^9 – 10^{10} cells (Del Monte, 2009; DeVita et al., 1975; Milo et al., 2010). The number of somatic mutations in a primary tumor cell (*M*) is the sum of mutations that were already present when the first oncogenic event occurred in the founder cell

(M_1) and the mutations that occurred subsequently during tumor evolution (M_2). Hence to obtain the mutation number of interest, M_2, we need to subtract M_1 from M. We get conservative estimates of M and M_1 from the WGS data as follows: M_1 is taken as the number of mutations in the tumor stem (clonal mutations, M_C) while M is taken as the average number of mutations in all subclones of the primary tumor; hence the clonal and subclonal mutation counts provide a measure of evolutionary time (Figure 5B). Excluding hypermutated cases, we find a median of $M_2 =$ 2300 mutations accumulated during the evolution of the primary tumor (Figure S5I). Mutations that are found at subclonal level fixed through cell divisions, and we assume a constant rate of mutation accumulation per division. The average somatic mutation rate was taken as 2.6×10^{-9} per base-pair and division (Milholland et al., 2017), which corresponds primarily to the clock-like mutation signature in our data (signature 1, Figure 1F). We observed a drop of the signature 1 contribution by 77% between clonal mutations (tumor stem) and late mutations (at the tips of the phylogenetic trees; Figure S5K), which implies an increase in mutation rate during tumor evolution provided that the clock-like rate is constant. Therefore, we accounted for the possibility of a fourfold increase of the somatic mutation rate during tumor evolution. To accumulate 2300 mutations with this range of mutation rates (2.6×10^{-9} to 10.6×10^{-9} mutations per base-pair and division), 66 to 268 cell divisions are required. To then reach a realistic tumor size of $10^9 – 10^{10}$ cells, most daughter cells (69-92%) must have died after cell division (Figures 5C and S5L). Hence, the combined balance of cell numbers and somatic mutations implies that massive cell death occurs during glioblastoma evolution and, on average, only 8-31% of cell divisions lead to tumor growth (Figure S5L).

The estimated high fraction of cell death is an average over the entire time course of evolution up to the resected primary tumor. Our WGS data show that a core set of pro-proliferative driver mutations (CNVs affecting *CDKN2A/B*, *PTEN*, or *EGFR*) occurred early during tumor evolution. *TERT* promoter mutations were also found in all but one primary tumor but their frequent subclonality indicates that they may occur a considerable time after tumor initiation. These mutations are not pro-proliferative but rather have been described as a key event that extends

cellular life span by healing short telomeres (Chiba et al., 2017). We modeled this scenario by assuming an initial growth regime with massive cell death before the *TERT* promoter mutation and, thereafter, a second regime with reduced cell death (Figure 5D). This model reproduced the observed distribution of tumors with clonal and subclonal *TERT* promoter mutations (focusing on the two canonical base pair substitutions in the *TERT* promoter (chr5:1295228 / chr5:1295250)). This fit was obtained when these mutations reduced the cell death rate by 6-28% (Figure 5E and S5N,O), corresponding to a selective advantage of 0.03 – 0.16. While this is an upper estimate because further mutations may have contributed to reducing cell death, the frequency of *TERT* promoter mutations is not matched by any other subclonal mutation, thus suggesting that other mutations are unlikely to have been major contributors. Notably, the selective advantage of *TERT* promoter mutations is larger than the average number for driver mutations (0.004 (Bozic et al., 2016)).

Finally, we estimated the time duration between the emergence of the founder cell and tumor diagnosis. To this end, we determined the average division rate of the tumor cells by dividing the number of mutations occurring after primary resection by the mutation rate and the time between primary and secondary resection. This yielded a lower bound of tumor cell division rate of approximately once in ten days. With the number of cell divisions given above that establish the primary tumor, we find that the founder cell emerges approximately two to seven years prior to diagnosis (Figure 5F). This range indicates that IDH^{WT} glioblastomas may undergo several years of evolution before being detected.

Discussion

The fast and incurable tumor recurrence in adult IDH^{WT} glioblastomas underlines the need for novel, more effective treatment approaches. To dissect the genetic basis of recurrence, we reconstructed the evolutionary histories of matched pairs of primary and recurrent IDH^{WT} glioblastomas, spanning the most common DNA methylome-based tumor subtypes. We found that the vast majority of recurrent tumors regrew from multiple genetic subclones in the primary tumors, thus reflecting the existing heterogeneity at the time of resection. This contrasts with the

368 previous suggestion of "clonal replacements" in the recurrent tumor under selective pressure
369 (Wang et al., 2016). However, in this previous study individual phylogenetic trees linking primary
370 and recurrent tumors were not derived. In our view, quantitative phylogenetic analysis is critical to
371 evaluate the effects of genetic drift (i.e., accumulation of neutral mutations) in the context of partial
372 sampling of the tumors. These two factors, drift and partial sampling, make related subclones in
373 primary and recurrent tumors appear to branch off from a common progenitor that was not
374 present in the actual samples. Nevertheless, the recurrent subclone is most likely a direct
375 descendant of the respective primary subclone, with the drawn samples just differing by neutral
376 mutations. Our data contrast with the description by Kim et al. (2015) of divergent glioblastoma
377 recurrences that share few genetic alterations with the primary tumor and originate from cells that
378 branched off early during tumorigenesis. This was associated with regrowth of the tumor in a
379 location distant from the original lesion. In our study, 90% of the tumors regrew in the same
380 location as the primary tumor. Therefore, the common pattern of clonal evolution identified here
381 may be characteristic of locally regrowing glioblastomas.
382 Our findings indicate that standard therapy exerted little selective pressure on most
383 recurrent tumors (with the possible exception of the few hypermutated cases). Consistent with this
384 idea, the vast majority of driver mutations were acquired prior to initial diagnosis and only few
385 drivers (in two of 21 cases: none) were acquired after initial treatment. The only gene recurrently
386 mutated in non-hypermutated relapse tumors was the non-coding RNA *XIST*, which has previously
387 been suggested as a driver in glioblastoma (Cheng et al., 2017; Yao et al., 2015). The paucity of
388 newly acquired driver genes in recurrent tumors contrasts with progressive evolution observed in
389 low grade glioma (Johnson et al., 2014), indicating that IDH^{WT} glioblastomas attained their full
390 aggressiveness already prior to initial diagnosis.
391 Combining data on genetic evolution and tumor growth, we inferred that IDH^{WT}
392 glioblastomas originated several years prior to diagnosis, suggesting that they undergo prolonged
393 evolution without causing clinical symptoms. Indeed, we uncovered a common path of early
394 tumorigenesis amongst our cohort: all tumors harbored at least one clonal copy number change in

EGFR (gain of chromosome 7), *CDKN2A/B* (loss of chromosome 9p or focal deletion) or *PTEN* (loss of chromosome 10/10q, with one tumor having a *PTEN* mutation instead of a loss). The observation that clonal 9p loss/focal deletion of *CDKN2A/B* always occurred in conjunction with chromosome 7 gain or 10/10q loss indicates that the former does not take place before the latter. Our inference of an early evolutionary origin of these chromosomal gains and losses is consistent with the overall abundance of these events in the respective TCGA cohort of 590 patients (Brennan et al., 2013; Grossman et al., 2016) and further studies (Sottoriva et al., 2013; Gerstung et al., 2017; Ozawa et al., 2014; Brastianos et al., 2017). It has also been suggested that *TERT* promoter mutations are very early events in IDH^{WT} glioblastomas (Barthel et al., 2018). We found that *TERT* promoter mutations, although acquired in the primary tumor, were subclonal in one third of our cohort. The subclonal resolution of our data places *TERT* promoter mutations at a subsequent stage of tumorigenesis, as has recently been suggested also for other tumors (Juratli et al., 2017, Landa et al., 2016). However, as *TERT* promoter mutations occur in all tumors (except for one primary sample), they appear to be associated with a selective advantage and, indeed, have reached clonality in two thirds of the tumors. During treatment, some (2/7) of the tumors with a subclonal VAF of the *TERT* promoter mutation progressed to clonality, but there was no overall increase in the respective VAFs that remained subclonal. Recent studies have linked *TERT* promoter mutation to better response to alkylating chemotherapy and longer patient survival in the subgroup of IDH^{WT} glioblastomas with *MGMT* promoter methylation (Arita et al., 2016; Nguyen et al., 2017), raising the interesting question of how chemotherapy may modulate the abundance of subclones depending on their mutation spectrum. Mutations in many non-coding RNAs were recurrent and often also clonal. Mutations in four of these, *HOTTIP*, *SNHG14*, *KCNQ1OT1* and *XIST*, have previously been implicated in cancer (Cheng et al., 2017; Jing et al., 2017; Wang et al., 2018; Yao et al., 2015). Mutations in *XIST* were also found in male samples and hence were not restricted to the inactive X chromosome (Jäger et al., 2013).

It is interesting to compare our finding that the mutated cell of origin of glioblastoma arose up to several years before diagnosis with reports of glioblastoma or high-grade glioma patients

who received an MRI scan because of new-onset seizures. In several instances, patients had mild abnormalities (hyperintense areas) or no abnormal findings in the MRI but then progressed to glioblastoma or high-grade glioma within several months (e.g., Nishi et al., 2009; Landy et al., 2000). Our estimate of the time of glioblastoma evolution relies on the count of somatic mutations and hence starts from the undetectable cell of origin. Comparing the number of cell divisions needed to accumulate the observed mutation counts with tumor size indicates a high rate of tumor cell death, so that only a minor fraction of cell divisions supports tumor growth. Therefore, it is plausible that the tumor remains of undetectable or barely detectable size for several years. To grow above the detection limit of current imaging techniques (about a millimeter in diameter), the tumor may need to acquire further mutations that stabilize cell survival. Activating *TERT* promoter mutations are a paradigmatic example of such mutations (Chiba et al., 2017), and our finding of subclonality of *TERT* promoter mutations in a sizeable fraction of tumors is consistent with the acquisition of this mutation in the process of tumor evolution.

Acknowledgments

The authors would like to thank the staff at the participating clinical centers of the German Glioma Network for their support; Stephan Wolf and the DKFZ Genomics and Proteomics Core Facility for their support in whole genome and RNA sequencing and for providing the Illumina Methylation Beadchips and related services; Jules Kerssemakers and the DKFZ Omics IT and Data Management Core Facility for their support in data processing and deposition at EGA. We thank Stefan Pusch and Daniel Schrimpf for support on processing and analysis of 450k/EPIC methylation data. We are grateful to Nagarajan Paramasivam and Naveed Ishaque for support on calling and interpretation of small mutations and structural variants. Roland Eils, Hai-Kun Liu, Markus Löffler and Stefan Pfister are acknowledged for discussions. This study was supported within the e:med program of the German Ministry of Education and Research (BMBF) by the collaborative research project 'SYS-GLIO – Systems-based prediction of the biological and clinical behavior of gliomas' (coordinator: P.L.; grant numbers 031A425). The German Glioma Network was funded from 2004 to 2012 by the German Cancer Aid (grant number 70-3163-Wi 3). A.Z. was supported by a Ph.D. grant from the Düsseldorf School of Oncology, Heinrich Heine University Düsseldorf, Germany. T.H. and V.K. are supported by the BMBF research project 'SYSMED-NB' (grant number 01ZX1307B). B.R. and P.L. thank the DKFZ-Heidelberg Center for Personalized Oncology (DKFZ-HIPO) for technical support and funding through HIPO_project H043.

Author contributions

V.K., G.R., T.H., and P.L., led the study and wrote the manuscript with input from all co-authors. V.K., M.S. and T.H. designed and supervised bioinformatic analyses; V.K., J.Y., P.B., Z.G., D.H. and M.F. performed bioinformatic analyses. V.K. and T.H. designed and performed mathematical modeling. Y.W. generated RNA and whole genome sequencing DNA sequencing data. D.S. and A.v.D. generated and evaluated DNA methylation data. A.Z., K.K., J.F., and G.R. generated and evaluated gene panel sequencing data. J.F. and G.R. performed the histological classification of tumor samples. J.F. and K.K. extracted nucleic acids and managed biobanking. D.J., B.R, and M.S. led and

performed RNA-seq and whole-genome sequencing experiments, and performed data analyses. K.L., J.C.T., C.H.-M. and GGN München, G.S. and GGN Dresden, M.Sa., B.H. and M.W. provided the patient samples and associated clinical details that made the study possible. D.J., B.R., and M.W. gave valuable input regarding study design, data analysis, and interpretation of results. G.R., T.H., and P.L. provided financial and technical infrastructure, and served as joint senior authors and project co-leaders. All authors read and approved the final manuscript.

Declaration of interests

G. Reifenberger has received research grants from Roche and Merck (EMD, Darmstadt), as well as honoraria for advisory boards from AbbVie. M. Weller has received research grants from Acceleron, Actelion, Bayer, Isarna, Merck, Sharp & Dohme, Merck (EMD, Darmstadt), Novocure, OGD2, Piqur and Roche, as well as honoraria for lectures or advisory board participation or consulting from BMS, Celldex, Immunocellular Therapeutics, Isarna, Magforce, Merck, Sharp & Dohme, Merck (EMD, Darmstadt), Northwest Biotherapeutics, Novocure, Pfizer, Roche, Teva and Tocagen. J. C. Tonn has received research honoraria for lectures or advisory board participation or consulting from Celldex, Roche, Siemens and BrainLab. The other authors declare no conflicts of interest.

References

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L. et al. (2013). Signatures of mutational processes in human cancer. Nature *22*, 415-421.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Arita, H., Yamasaki, K., Matsushita, Y., Nakamura, T., Shimokawa, A., Takami, H., Tanaka, S., Mukasa, A., Shirahata, M., Shimizu, S. et al. (2016). A combination of TERT promoter mutation and MGMT methylation status predicts clinically relevant subgroups of newly diagnosed glioblastomas. Acta Neuropathol. Commun. *4*, 79.

Aryee M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. Bioinformatics *30*, 1363-1369.

Barthel, F.P., Wesseling, P., and Verhaak, R.G.W. (2018). Reconstructing the molecular life history of gliomas. Acta Neropathol. *135*, 649-670.

Becht, E., Giraldo, N.A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W.H. et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biology *17*, 218.

Bozic, I., Gerold, J.M., and Nowak, M.A. (2016). Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. PLoS Comput. Biol. 12, e1004731.

Brastianos, P.K., Nayyar, N., Rosebrock, D., Leshchiner, I., Gill, C.M., Livitz, D., Bertalan, M.S., D'Andrea, M., Hoang, K, Aquilanti, E. et al. (2017). Resolving the phylogenetic origin of glioblastoma via multifocal genomic analysis of pre-treatment and treatment-resistant autopsy specimens. NPJ Precision Oncology *33*, 1-33.

Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H. et al. (2013). The somatic genomic landscape of glioblastoma. Cell *155*, 462-477.

Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L, Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature *22*, 469-474.

Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. Biometrika *95*, 759-771.

Cheng, Z., Li, Z., Ma, K., Li, X., Tian, N., Duan, J., Xiao, X., and Wang, Y. (2017). Long Non-coding RNA XIST Promotes Glioma Tumorigenicity and Angiogenesis by Acting as a Molecular Sponge of miR-429. J. Cancer *8*, 4106-4116.

Chiba, K., Lorbeer, F.K., Hunter Shain, A., McSwiggen, D.T., Schruf, E., Oh, A., Ryu, J., Darzacq, X., Bastian, B.C., and Hockemeyer, D. (2017). Mutations in the promoter of the telomerase gene *TERT* contribute to tumorigenesis by a two-step mechanism. Science *357*, 1416-1420.

deCarvalho, A.C., Kim, H., Poisson, L.M., Winn, M.E., Mueller, C., Cherba, D., Koeman, J., Seth S., Protopopov, A., Felicella, M., et al. (2018). Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat. Genet. *50*, 708-717.

Del Monte, U. (2009). Does the cell number 10(9) still really fit one gram of tumor tissue? Cell Cycle *8*, 505-506.

DeVita, V.T.Jr., Young, R.C., and Canellos, G.P. (1975). Combination versus single agent chemotherapy: a review of the basis for selection of drug treatment of cancer. Cancer *35*, 98-110.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Drachman, D. (2005). Do we have brain to spare? Neurology *64*, 2004-2005.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available from: https://www.R-project.org/.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E, Ponting, L. et al. (2016). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777-D783.

Francis, J.M., Zhang, C.Z., Maire, C.L., Jung, J., Manzo, V.E., Adalsteinsson, V.A., Homer, H, Haidar, S., Blumenstiel, B., Pedamallu, C.S. et al. (2014). *EGFR* variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. Cancer Discov. *4*, 956-971.

Gerlinger, M., and Swanton, C. (2010). How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. BJC *103*, 1139-1143.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Mitchell, T.J., Rubanova, Y., Anur, P., Rosebrock, D., Yu, K. et al. (Preprint from September 20, 2017). The evolutionary history of 2,658 cancers. BioRxiv *161562*. Available from: https://doi.org/10.1101/161562.

Giessler, K.M., Kleinheinz, K., Huebschmann, D., Balasubramanian, G.P., Dubash, T.D., Dieter S.M., Siegl, C., Herbst, F., Weber, S., Hoffmann, C.M. et al. (2017) Genetic subclone architecture of tumor clone-initiating cells in colorectal cancer. J. Exp. Med. *214*, 2073-2088.

Goldberg-Zimring, D., Talos, I.F., Bhagwat, J.G., Haker, S.J., Black, P.M., and Zou, K.H. (2005). Statistical validation of brain tumor shape approximation via spherical harmonics for image-guided neurosurgery. Acad. Radiol. *12*, 459-466.

Gonzalez-Perez A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer. Nat. Methods *10*, 1081–1082.

Greaves, M., Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306-313.

Gröbner, S.N., Worst, B.C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V.A., Johann, P.D., Balasubramanian, G.P., Segura-Wang M., Brabetz S. et al. (2018). The landscape of genomic alterations across childhood cancers. Nature *555*, 321-327.

Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a Shared Vision for Cancer Genomic Data. N. Engl. J. Med. *375*, 1109-1112.

Hegi, M.E., Diserens A.C., Gorlia T., Hamou M.F., de Tribolet N., Weller M., Kros J.M., Hainfellner J.A., Mason W., Mariani L., et al. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. N. Engl. J. Med. *352*:997-1003.

Herculaono-Houzel, S., Mota, B., and Lent, R. (2006). Cellular scaling rules for rodent brains. Proc. Natl. Acad. Sci. USA *103*, 12138-12143.

Hu, L.S., Ning, S., Eschbacher, J.M., Baxter, L.C., Gaw, N., Ranjbar, S., Plasencia, J., Dueck, A.C., Peng, S., Smith, K.A. et al. (2017). Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. Neuro Oncol. *19*, 128-137.

Huebschmann, D., Gu, Z., and Schlesner, M. (2015). YAPSA: Yet Another Package for Signature Analysis. Available from: https://bioconductor.org/packages/devel/bioc/html/YAPSA.html.

Jaffe A.E. (2018). FlowSorted.Blood.450k: Illumina HumanMethylation data on sorted blood cell populations. Available from: https://www.bioconductor.org/packages/release/data/experiment/html/FlowSorted.Blood.450k.html.

Jaffe A.E., and Kaminsky Z.A. (2018). FlowSorted.DLPFC.450k: Illumina HumanMethylation data on sorted frontal cortex cell populations. Available from: https://bioconductor.org/packages/release/data/experiment/html/FlowSorted.DLPFC.450k.html.

Jäger, N., Schlesner, M., Jones, D.T., Raffel, S., Mallm, J.P., Junge, K.M., Weichenhan, D., Bauer, T., Ishaque, N., Kool, M. et al. (2013). Hypermutation of the inactive X chromosome is a frequent event in cancer. Cell *155*, 567-581.

Jin, N., Yang, L.Y., and Xu, Z.P. (2017). Long non-coding RNA HOTTIP is able to predict poor prognosis in various neoplasms: A meta-analysis. Mol. Clin. Oncol. *7*, 263-266.

Johnson, B.E., Mazor, T., Hong, C., Barnes, M., Aihara, K., McLean, C.Y., Fouse, S.D., Yamamoto, S., Ueda, H., Tatsuno, K. et al. (2014). Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. Science *343*, 189-193.

Juratli, T.A., Thiede, C., Koerner, M.V.A., Tummala, S.S., Daubner, D., Shankar, G.M., Williams, E.A., Martinez-Lage, M., Soucek, S., Robel, K. et al. (2017). Intratumoral heterogeneity and TERT promoter mutations in progressive/higher-grade meningiomas. Oncotarget *8*, 109228-109237.

Khan, A., Zhang, X. (2016). dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Res., *44*, D164–D171.

Kim, H., Zheng, S., Amini, S.S., Virk, S.M., Mikkelsen, T., Brat, D.J., Grimsby, J., Sougnez, C., Muller, F., Hu, J. et al. (2015). Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. Genome Res. *3*, 316-27.

Kimura M. (1969). The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. Genetics *61*, 893-903.

Kleinheinz, K., Bludau, I., Hübschmann, D., Heinold, M., Kensche, P., Gu, Z., López, C., Hummel, M., Klapper, W., Möller, P. et al. (Preprint from October 29, 2017). ACEseq-allele specific copy number estimation from whole genome sequencing. BioRxiv *210807*. Available from: https://doi.org/10.1101/210807.

Landa, I., Ibrahimpasic, T., Boucai, L., Sinha, R., Knauf, J.A., Shah, R.H., Dogan, S., Ricarte-Filho, J.C., Krishnamoorthy, G.P., Xu, B. et al. (2016). Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers. J. Clin. Invest. *126*, 1052-1066.

Landy, H.J., Lee, T.T., Potter, P., Feun, L. and Markoe, A. (2000). Early MRI findings in high grade glioma. J Neurooncol. *47*, 65-72.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. Proc. Natl. Acad. Sci. USA *112*, E6496-E6505.

Love, M.I., Huber, W., Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, *15*, 550.

614 Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavanee, W.K.,
615 Ohgaki, H., Wiestler, O.D., Kleihues, P., and Ellison, D.W. (2016). The 2016 World Health
616 Organization Classification of Tumors of the Central Nervous System: a summary. Acta
617 Neuropathol. *131*, 803-820.

618 Meyer, M., Reimand, J., Lan, X., Head, R, Zhu, X., Kushida, M., Bayani, J., Pressey, J.C., Lionel, A.C.,
619 Clarke, I.D. et al. (2013). Single-cell derived clonal analysis of human glioblastoma links
620 functional and genomic heterogeneity. Proc. Natl. Acad. Sci. *112*, 851-856.

621 Milholland, B., Dong, X., Zhang, L, Hao, X., Suh, Y., and Vijg, J. (2017). Differences between
622 germline and somatic mutation rates in humans and mice. Nat. Commun. *8*, 15183.

623 Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers—the
624 database of key numbers in molecular and cell biology. Nucleic Acids Res. *38*, D750-D753.

625 Nathanson, D.A., Gini, B., Mottahedeh, J., Visnyei, K., Koga, T., Gomez, G., Eskin, A., Hwang, K.,
626 Wang, J., Masui, K. et al. (2014). Targeted Therapy Resistance Mediated by Dynamic Regulation
627 of Extrachromosomal Mutant EGFR DNA. Science *343*, 72-76.

628 Nishi, N., Kawai, S., Yonezawa, T., Fujimoto, K., and Masui, K. (2009). Early Appearance of High
629 Grade Glioma on Magnetic Resonance Imaging. Neurol. Med. Chir. (Tokyo) *49*, 8-12.

630 Nguyen, H.N., Lie, A., Li, T., Chowdhury, R., Liu, F., Ozer, B., Wei, B., Green, R.M., Ellingson B.M.,
631 Wang, H.J. et al. (2017). Human TERT promoter mutation enables survival advantage from
632 MGMT promoter methylation in IDH1 wild-type primary glioblastoma treated by standard
633 chemoradiotherapy. Neuro Oncol. *19*:394-404

634 Ostrom, Q.T., Gittleman, H., Xu, J., Kromer, C., Wolinsky, Y., Kruchko, C., and Barnholtz-Sloan, J.S.
635 (2016). Primary Brain and Other Central Nervous System Tumors Diagnosed in the United
636 States in 2009–2013. Neuro Oncol. *18*, v1-v75.

637 Ozawa, T., Riester, M., Cheng, Y.-K., Huse, J.T., Squatrito, M., Helmy, K., Charles, N., Michor, F., and
638 Holland, E.C. (2014). Most human non-GCIMP glioblastoma subtypes evolve from a common
639 proneural-like precursor glioma. Cancer Cell *26*, 288-300.

640 Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016).
641 OncodriveFML: a general framework to identify coding and non-coding regions with cancer
642 driver mutations. Genome Biology *17*, 128.

643 Pakkenberg, B., Gundersen, H.J., Mortensen, E.L., Lauritzen, M.J., Jeune, B., Regeur, L., West, M.J.,
644 and Schwartz, T.W. (1997). The normal brain: a new knowledge in different fields. Ugeskr Laeger
645 *159*, 723-727.

646 Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P.,
647 Nahed, B.V., Curry, W.T., Martuza, R.L. et al. (2014). Single-cell RNA-seq highlights intratumoral
648 heterogeneity in primary glioblastoma. Science *344*, 1396-1401.

649 Qazi, M.A., Vora, P., Venugopal, C., Sidhu, S.S., Moffat, J., Swanton, C., and Singh, S.K. (2017).
650 Intratumoral heterogeneity: pathways to treatment resistance and relapse in human
651 glioblastoma. Ann. Oncol. *28*, 1448-1456.

652 Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., WGS500 Consortium, Wilkie, A.O.M.,
653 McVean, G., and Lunter G. (2014). Integrating mapping-, assembly- and haplotype-based
654 approaches for calling variants in clinical sequencing applications. Nat. Genet. *46*, 1–9.

655 Rubio-Perez, C., Tamborero, D., Schroeder, M.P., Antolín, A.A., Deu-Pons, J., Perez-Llamas, C.,
656 Mestres, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2015). In silico prescription of anticancer
657 drugs to cohorts of 28 tumor types reveals novel targeting opportunities. Cancer Cell *27*, 382-
658 396.

659 Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001).
660 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

661 Snuderl, M., Fazlollahi, L., Le, L.P., Nitta, M., Zhelyazkova, B.H., Davidson, C.J., Akhavanfard, S.,
662 Cahill, D.P., Aldape, K.D., Betensky R.A. et al. (2011). Mosaic Amplification of Multiple Receptor
663 Tyrosine Kinase Genes in Glioblastoma. Cancer Cell *20*, 810-817.

664 Sottoriva, A., Spiteri, I., Piccirillo, S.G.M., Touloumis, A. Collins, V.P., Marioni, J.C., Curtis, C., Watts,
665 C., and Tavaré, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer
666 evolutionary dynamics. Proc. Natl. Acad. Sci.*110*, 4009-4014.

667 Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D.A., Jones, D.T., Konermann, C., Pfaff, E.,
668 Tönjes, M., Sill, M., Bender, S. et al. (2012). Hotspot mutations in H3F3A and IDH1 define distinct
669 epigenetic and biological subgroups of glioblastoma. Cancer cell *22*, 425-437.

670 Tischler, G., and Leonard, S. (2014). biobambam: tools for read pair collation based algorithms
671 on BAM files. Source Code Biol. Med. *9*, 1-18.

672 Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D.I., Zairis, S., Abate, F., Liu, Z., Elliott,
673 O., Shin, Y.J. et al. (2016). Clonal evolution of glioblastoma under therapy. Nat. Genet. *48*, 768-
674 776.

675 Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic
676 variants from high-throughput sequencing data. Nucleic Acids Res. *38*, 1–7.

677 Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., deCarvalho, A.C., Lyu, S., Li, P., Li, Y.,
678 et al. (2017). Tumor evolution of glioma-intrinsic gene expression subtypes associates with
679 immunological changes in the microenvironment. Cancer Cell *32*, 42-56.

680 Wang, Q., Teng, Y., Wang, R., Deng, D., You, Y., Peng, Y., Shao, N., and Zhi, F. (2018). The long
681 non-coding RNA SNHG14 inhibits cell proliferation and invasion and promotes apoptosis by
682 sponging miR-92a-3p in glioma. Oncotarget *9*, 12112-12124.

683 Weller, M., van den Bent, M., Tonn, J.C., Stupp, R., Preusser, M., Cohen-Jonathan-Moyal, E.,
684 Henriksson, R., Le Rhun, E., Balana, C., Chinot, O. et al. (2017). European Association for Neuro-
685 Oncology (EANO) guideline on the diagnosis and treatment of adult astrocytic and
686 oligodendroglial gliomas. Lancet Oncol. *18*, e315-329.

687 Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A., and Sottoriva, A. (2016). Identification of
688 neutral tumor evolution across cancer types. Nat. Genet. *48*, 238-244.

689 Yao, Y., Ma, J., Xue, Y., Wang, P., Li, Z., Liu, J., Chen, L., Xi, Z., Teng, H., Wang, Z. et al. (2015).
690 Knockdown of long non-coding RNA XIST exerts tumor-suppressive functions in human
691 glioblastoma stem cells by up-regulating miR-152. Cancer Lett. *359*, 75-86.

692 Yates, L.R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T.,
693 Alexandrov, L.B., Larsimont, D., Davies, H. et al. (2015). Subclonal diversification of primary
694 breast cancer revealed by multiregion sequencing. Nat. Med. *21*, 751-760.

695 Zacher, A., Kaulich, K., Stepanow, S., Wolter, M., Köhrer, K., Felsberg, J., Malzkorn, B., and
696 Reifenberger, G. (2017). Molecular Diagnostics of Gliomas Using Next Generation Sequencing
697 of a Glioma-Tailored Gene Panel. Brain Pathol. *27*, 146-159.
698

699

700

Figure legends

Figure 1. Clinical data and mutation statistics. See also Figure S1.

(A) Sampling strategy.

(B) Overview of clinical parameters and sequencing approaches (see also Figure S1).

(C) Number of SNVs and small insertions/deletions called in primary and recurrent samples of the discovery set (red lines, median; red dots, hypermutated cases).

(D) Numbers of common and private SNVs and small insertions/deletions of the discovery set (red lines, median; red dots, hypermutated cases).

(E) Relative numbers corresponding to (D) per tumor.

(F-H) Mutational signatures of shared (F) and private SNVs (G) from non-hypermutated cases of the discovery set, as well as mutational signatures of hypermutated cases (H). Data are represented as mean ± SEM.

Figure 2. Recurrent driver mutations in paired primary and recurrent tumor samples from 21 glioblastoma patients. See also Figure S2.

(A) Mutational spectrum in likely driver genes. Coding genes are shown if targeted by a structural variation, SNV, indel, homozygous deletion or a high-level amplification in at least two patients (excluding hypermutated samples; mutations in *MSH6* are additionally shown due to their correlation with a hypermutation genotype). Non-coding genes are shown if targeted by a SNV or an insertion/deletion in more than five patients. *TERT* promoter mutations are shown at the bottom. Primary and recurrent samples of a tumor are shown in pairs and separated by vertical lines (left primary, right secondary). Different colors mark different types of mutations while the color code at the bottom indicates DNA methylation subtypes.

(B) Chromosomal gains and losses in primary and relapsed tumors.

(C) Inter-chromosomal and intra-chromosomal structural variants in primary and relapsed tumors. Translocation partners are marked if targeting a gene or the vicinity of a super-enhancer (based on

dbSUPER; Khan et al., 2016). Genes targeted by intra-chromosomal variants are highlighted if they were present in the driver genes list. Numbers in brackets indicate the number of recurrences of a structural variant (tr., inter-chromosomal translocation; ic., intra-chromosomal variant).
(D, E) Two examples of ongoing evolution on chromosome 7. Both cases can be explained by an initial gain of the chromosome, followed by partial losses in the recurrent (D) or the primary tumor (E) (black, normal copy number; green, gains; red, losses; light blue lines, number of A- and B-alleles).

Figure 3. Phylogenetic inference. See also Figures S3 and S4.

(A) Variant allele frequencies at loci with normal copy numbers (excluding hypermutated samples and, for clarity, showing mutations in functional regions only). Clonal and subclonal mutations were classified based on a binomial test for clonality (significance level $\alpha = 0.05$).
(B) Principle of phylogenetic inference from whole-genome sequencing data, with fractions of mutated and reference reads at mutated loci (left panel), the inferred subclonal distribution (middle panel) and the resulting phylogenetic tree (right panel). The mutated DNA fraction within tumor subclones is indicated by a darker shading; stars mark examples of loci with copy number changes. The data shown are for the two subclones of the recurrent tumor in (H).
(C, D) Summary statistics on inferred phylogenies, with number of subclones (C) and their relative sizes (D).
(E) Layout of phylogenetic trees. Vertical branch lengths scale with the number of mutations, circle areas scale with relative subclone sizes.
(F,G) Examples of phylogenetic trees with oligoclonal origin of the recurrent tumor, with (F) and without (G) recurrent driver mutations acquired after primary resection (yellow, primary tumor; orange, recurrent tumor; blue branches, common stem; grey branches, mutations originating in primary tumor; green branches, mutations found only in relapse tumor; asterisk marks convergent evolution; mutations in driver genes and copy number changes of chromosomes 7, 9 and 10 are indicated).

(H) Monoclonal origin of the recurrent tumor; primary and recurrent tumors are separated by a long branch and several new driver mutations (mutations in driver genes and copy number changes of chromosomes 7, 9 and 10 are indicated).
(I) Origin of the recurrent tumor relative to origin of the primary tumor (excluding hypermutated cases). The distance between the most recent common ancestors (MRCA) of primary and recurrent samples (measured in mutation counts) is provided as a measure for tumor origin.
(K) Consensus tree of all sample pairs. Branch widths and font sizes scale with the number of cases supporting a connection and mutation, respectively. The median number of clonal mutations and the median of the maximal number of mutations per subclone are indicated for primary and recurrent tumors. Driver mutations are indicated at particular tree branches [distinguishing clonal, subclonal, present in primary or recurrence only] if they are found there in at least two tumors and are frequent overall [present at any position in at least three (coding genes) or five (non-coding genes) non-hypermutated tumors].
All data are from the discovery set subjected to deep WGS.

Figure 4. Patterns of mutations during different stages of tumorigenesis. See also Figure S5.
(A) Number of clonal and subclonal copy number changes on Chr. 7, 9 and 10.
(B) Number of clonal and subclonal mutations (SNVs and indels) in driver genes that were clonal in at least one tumor. Red bars represent the numbers of tumors harboring at least one of the listed driver mutations clonally. In a few cases, mutations could not unambiguously be placed on the phylogenetic trees as clonal or subclonal; these were counted as clonal when the probability of placing them in the tumor stem was >10%.
(C) Canonical (red) and non-canonical mutations (grey) in the *TERT* promoter region.
All data are from the discovery set subjected to deep WGS.

Figure 5. Evolutionary dynamics during glioblastoma growth. See also Figure S5.

781 (A) Model of tumor growth. Cells divide at rate λ, die at rate δ and accumulate mutations at rate $\lambda\mu$.
782 Clinically detectable tumors have cell counts in the order of 10^9-10^{10}, each of which on average
783 accumulated 2300 mutations between tumor initiation and clinical detection (median of 17 non-
784 hypermutated tumors).
785 (B) Schematic of the mutation counts used for rate estimation. Tumors are initiated by a somatic
786 cell with M_1 prior (neutral) mutations. Extant subclonal diversification occurs after M_C mutations
787 (i.e., at the most recent common ancestor of the tumor). Individual cells accumulate additional
788 mutations during tumor growth (resulting in a total of *M* mutations per cell at clinical
789 presentation).
790 (C) The numbers of cell divisions and mutations scale logarithmically with the number of tumor
791 cells. The slope of the growth curve is determined by the rate of cell death. Hence knowledge of
792 tumor cell number and number of cell divisions allows to infer the rate of cell death.
793 (D) Model for the acquisition of canonical *TERT* promoter mutations during glioblastoma evolution.
794 A transformed cell divides due to initial pro-proliferative mutations (including the common CNVs in
795 the tumor stem). After capturing an additional *TERT* promoter mutation with probability μ dt, the
796 cellular death rate is reduced to δ_{TERT}.
797 (E) Inferred and modeled cumulative distribution of the tumor fraction with canonical *TERT*
798 promoter mutation (estimated from its most likely position in the phylogenetic trees of the 19
799 cases with canonical *TERT* promoter mutation).
800 (F) Model of glioblastoma growth (individual subclones are labeled by different shades of yellow to
801 red for clearer distinction; MRCA, most recent common ancestor).
802
803
804
805

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Peter Lichter (peter.lichter@dkfz.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patients and tissue samples

We collected matched tissue samples from the initial surgery before treatment and a second surgery for recurrent tumor growth from 50 patients with IDH^{WT} glioblastoma, World Health Organization (WHO) grade IV. The patients were identified in the central database of the German Glioma Network (GGN) or the database of the Central Nervous System (CNS) tumor tissue bank at the Department of Neuropathology, Heinrich Heine University, Düsseldorf, Germany. The GGN is a prospective, non-interventional cohort study involving eight clinical centers at University Hospitals in Germany (www.gliomnetzwerk.de), and was supported by the German Cancer Aid from 2004 to 2012. Patients provided their written informed consent for participating in the GGN and the use of their tissue samples and clinical data for research purposes according to the research proposals approved by the Institutional Review Boards of the participating institutions. The present study was approved by the institutional review board of the Medical Faculty, Heinrich Heine University, Düsseldorf, Germany (study number 4940). Histology of all tumors was centrally reviewed and confirmed to correspond to glioblastoma, IDH^{WT}, World Health Organization (WHO) grade IV according to the WHO classification of central nervous system tumors 2016 (Louis et al., 2016). All tissue specimens used in the present study for extraction of nucleic acids and molecular analyses were histologically evaluated to assure that they consisted of vital tumor tissue. Tissue samples corresponding mostly to necrosis or reactive changes were excluded. Relevant clinical data retrieved for each patient included information on age at diagnosis, gender, tumor location, extent

of resection, postoperative therapy, interval between primary and secondary surgery, and overall survival. Table S1 shows a summary of the respective patient data.

METHOD DETAILS

Immunohistochemical stainings

Immunohistochemical stainings for ATRX (alpha-thalassemia/mental retardation syndrome, X-linked) protein expression were performed on representative formalin-fixed and paraffin-embedded tissue sections from selected cases of primary and recurrent tumors using an automated immunostainer (Dako, Kopenhagen, Denmark) and the UltraVision™ Quanto horseradish peroxidase detection system with 3,3'-diaminobenzidine tetrahydrochloride as chromogen (Thermo Fisher Scientific, Freemont, CA). Sections were pretreated by heating them in retrieval solution (Dako) at pH 9.0 for 20 min. As primary anti-ATRX antibody, we used the mouse monoclonal IgG1 clone AX1 diluted 1:200 as recommended by the manufacturer (Dianova, Hamburg, Germany). All immunohistochemical sections were counterstained with hematoxylin.

Extraction of DNA and RNA

DNA was extracted from deep-frozen tumor tissue samples from 35 patients, including all patients whose tissue samples were subjected to WGS with the JETQUICK Tissue DNA Spin Kit (Genomed, Loehne, Germany). Extraction of constitutional (leukocyte) DNA from frozen peripheral blood samples was performed with the Pure Link Genomic DNA Mini Kit (Invitrogen, Carlsbad, CA). The RNeasy Mini Kit (Qiagen, Hilden, Germany) was used for extraction of total RNA from deep-frozen tissue samples. The GeneRead DNA FFPE Kit (Qiagen) was used for DNA extraction from formalin-fixed and paraffin-embedded (FFPE) samples of 15 patients whose tumor samples were subjected to panel sequencing. All nucleic acid extractions were carried out according to the respective kit manufacturer's protocol. Quality of extracted DNA was assured by spectrophotometric analysis and agarose gel electrophoresis. RNA quality was determined on an Agilent 2100 Bioanalyzer (Agilent

Technologies, Santa Clara, CA). Only samples showing a RNA integrity number (RIN) of 6.5 or more were used for RNA sequencing.

DNA and RNA library preparation

DNA libraries were prepared according to the Illumina TruSeq Nano DNA Library protocol using the TruSeq DNA Nano kit (Illumina, Hayward, CA) and sequenced on HiSeq X using the HiSeq X Ten Reagent Kit v2.5 (both Illumina, Hayward, CA). Total, strand-specific RNA libraries were prepared according to the Illumina TruSeq Stranded Total RNA Sample Preparation Guide (Illumina, Hayward, CA), with Ribo Zero Gold (Epicentre, Madison, WI) depletion of rRNA.

Gene panel next generation sequencing

In total, we analyzed pairs of primary and recurrent IDH^{WT} glioblastomas from 43 patients (including 14 of the 21 patients whose tumors were analyzed by whole genome sequencing) using next generation sequencing of a gene panel covering 50 selected genes. Libraries for gene panel sequencing were prepared using the Ion AmpliSeq™ Library 2.0 Kit (Life Technologies, Carlsbad, CA, USA) and two customized AmpliSeq™ gene panels. Gene panel 1 has been published elsewhere (Zacher et al., 2017) and covers the complete coding sequences (cds) or selected genomic regions including mutation hot spots (region) of the following 20 genes: *NRAS* (cds), *FUBP1* (cds), *CDKN2C* (cds), *H3F3A* (region), *IDH1* (region), *PIK3CA* (cds), *TERT* (region), *PIK3R1* (cds), *EGFR* (cds), *BRAF* (region), *CDKN2A* (cds), *CDKN2B* (cds), *PTEN* (cds), *RB1* (cds), *IDH2* (region), *TP53* (cds), *NF1* (cds), *CIC* (cds), *NF2* (cds) and *ATRX* (cds). Gene panel 2 covers complete cds or regions including mutation hot spots of the following 30 genes: *MDM4* (region), *PIK3C2B* (cds), *ACVR1* (region), *MSH6* (cds), *MSH2* (cds), *RAB7A* (region), *MLH1* (cds), *PDGFRA* (cds), *MYB* (region), *CCND3* (region), *DAXX* (cds), *HIST1H3B* (cds), *CDK6* (region), *MET* (region), *PMS2* (cds), *MYBL1* (region), *FGFR1* (region), *NTRK2* (cds), *TSC1* (cds), *MTAP* (cds), *CCND1* (region), *CDK4* (region), *PTPN11* (cds), *MDM2* (region), *CCND2* (region), *KRAS* (cds), *IRS2* (region), *GABRG3* (region), *TSC2* (cds) and *PPM1D* (cds). Library preparation for gene panel sequencing was carried out according to the manufacturer's

protocol (Zacher et al., 2017) (MAN 006735 Rev 5.0, Life Technologies). A total of four primer pools were PCR amplified using 10 ng of genomic DNA per primer pool as template. Emulsion PCR and enrichment was carried out on the Ion OneTouch™ 2 System using the Ion PI™ Template OT2 200 Kit v3 (MAN0009133 Rev A, Life Technologies). The Ion Proton™ System (Life Technologies) was used for sequencing using the Ion PI™v2 chips and the Ion PI™ 200 Sequencing 200 Kit v3 (MAN0009136 Rev.A, Life Technologies). Overall, at least 1 million reads were generated per library. Data analysis was performed with the Torrent Suite™ software 4.4 employing the implemented TMAP algorithm (Life Technologies). Variants were detected using the VariantCaller Plugin v4.4 and the predefined parameter set 'somatic high stringency' (Life Technologies, for detailed annotation information see Zacher et al., 2017). The Integrative Genomics Viewer (IGV) software was used to visualize the read alignment and check for possible errors. Copy number variations were identified with the NextGENe® v.2.3.4 software (SoftGenetics, State College, PA).

Alignment of whole genome sequencing reads

Read pairs were mapped to the human reference genome (build 37, version hs37d5) using bwa mem (Li and Durbin, 2009) (version 0.7.8, with minimum base quality threshold zero [-T 0] and remaining settings left at default values), followed by coordinate-sorting with bamsort (with compression option set to fast (1)) and marking duplicate read pairs with bammarkduplicates (with compression option set to (9)); both are part of biobambam (Tischler and Leonard, 2014) package version 0.0.148.

Detection of SNVs and indels

Somatic SNVs and indels in matched tumor normal pairs were identified using the DKFZ core variant calling workflows of the ICGC PCAWG project (https://dockstore.org/containers/quay.io/pancancer/pcawg-dkfz-workflow). Briefly, in the DKFZ SNV pipeline candidate SNV calls were generated by samtools (Li et al., 2009) and bcftools (version 0.1.19), and potential variants called in the tumor were followed by a lookup of the corresponding

positions in the control. To enable calling of variants with low allele frequency we disabled the Bayesian model (by setting -p 2). Thus, all positions containing at least one high quality non-reference base are reported as candidate variant. The resulting raw calls were categorized into putative somatic variants and others (artifacts, germline) based on the presence of variant reads in the matched normal sample. The frequency of all putative somatic variants was then refined by checking for potential redundant information due to overlapping reads, and precise base counts for each strand were determined. All variants were annotated with dbSNP141, 1000 Genomes (phase 1), Gencode Mapability track, UCSC High Seq Depth track, UCSC Simple-Tandemrepeats, UCSC Repeat-Masker, DUKE-Excluded, DAC-Blacklist, UCSC Selfchain. The confidence for each variant was then determined by a heuristic scoring scheme taking the aforementioned tracks into account. In addition, variants with strong read biases which fell into a bias-prone context according to the strand bias filter were removed. High confidence variants (confidence score 8-10) were reported. For indel calling, Platypus (Rimmer et al., 2014) version 0.7.4 was used. All candidate indel variants were categorized into putative somatic and other based on the genotype likelihoods (matched genotype 0/0 for somatic indels). High confidence somatic variants were required to either have the Platypus filter flag PASS or pass custom filters allowing for low variant frequency using a scoring scheme. Candidates with the badReads flag, alleleBias, or strandBias were discarded if the variant allele frequency was <10%. Additionally, combinations of Platypus non-PASS filter flags, bad quality values, low genotype quality, very low variant counts in the tumor, and presence of variant reads in the control were not tolerated. SNVs and indels were annotated using ANNOVAR (Wang et al., 2010) according to GENCODE gene annotation (version 19) and overlapped with variants from dbSNP (Sherry et al., 2001) (build 141) and the 1000 Genomes Project database.

Detection of structural variants (SVs)

Genomic structural rearrangements were detected using SOPHIA v.34. Briefly, SOPHIA uses supplementary alignments as produced by bwa mem as indicators of a possible underlying SV. SV candidates are filtered by comparing them to a background control set of sequencing data

(obtained using normal blood samples from a background population database of 3261 patients from published TCGA and ICGC studies and both published and unpublished DKFZ studies, sequenced using Illumina HiSeq 2000, 2500 (100 bp) and HiSeq X (151 bp) platforms and aligned uniformly using the same workflow as in this study). An SV candidate is discarded if (i) it has more than 85% of read support from low quality reads; (ii) the second breakpoint of the SV was unmappable in the sample and the first breakpoint was detected in 10 or more background control samples; (iii) an SV with two identified breakpoints had one breakpoint present in at least 98 control samples (3% of the control samples); or (iv) both breakpoints have less than 5% read support. SVs were annotated as somatic if the respective event had no support in the matched normal sample.

Detection of copy number aberrations

Allele-specific copy-number aberrations were detected using ACEseq (allele-specific copy-number estimation from whole genome sequencing) (Kleinheinz et al., 2017). ACEseq determines absolute allele-specific copy numbers as well as tumor ploidy and tumor cell content based on coverage ratios of tumor and control as well as the B-allele frequency (BAF) of heterozygous single nucleotide polymorphisms (SNPs). SVs called by SOPHIA were incorporated to improve genome segmentation. The estimated tumor cell content was compared to the doubled median of the somatic SNV mutation allele frequency distribution and adjusted if the estimates deviated by more than 10% from each other. Genomic segments were annotated as losses if their total copy number (TCN) was smaller than the sample's ploidy by at least 0.3. Segments with an estimated TCN < 0.5 were annotated as homozygous deletions. Analogous to losses, gains were identified as segments with ploidy – TCN < -0.3. High level amplifications were classified as segments with TCN > 2.5 x ploidy. Loss of heterozygosity was defined as segments with the copy number of the minor allele < 0.3.

Alignment and counting of RNA sequencing reads

Sequencing reads were aligned to the human genome (hg19 build) with the Gencode reference transcriptome (v19) using STAR (v2.3.0e) (Dobin et al., 2013). Read counts for each gene were determined as the total number of reads mapping to exons using htseq-count (0.6.0) (Anders et al., 2015) and the GENCODE transcript model (v19).

450k/EPIC methylation arrays

DNA methylation profiling was performed by using 450k or EPIC methylation bead arrays (Illumina, Hayward, CA) according to standard protocols. Data were processed and analysed as reported elsewhere (Capper et al., 2018).

Signature analysis

Supervised mutational signature analysis was performed with the R package YAPSA (Huebschmann et al., 2015). Using the function LCD_complex_cutoff(), we computed a non-negative least squares (NNLS) decomposition of the mutational catalogue with 30 known signatures from COSMIC (http://cancer.sanger.ac.uk/cosmic/signatures)). In order to increase specificity, LCD_complex_cutoff() applies the NNLS algorithm twice. A first NNLS is run proposing all supplied signatures to the decomposition, then a second NNLS is run again with a reduced set of signatures consisting only of those signatures whose exposures were higher than a certain signature-specific cutoff. The signature-specific cutoffs were determined in a random operator characteristic (ROC) analysis using publicly available data on mutational catalogues of 7,042 cancer samples (507 from whole genome sequencing and 6,535 from whole exome sequencing; Alexandrov et al., 2013) and mutational signatures (http://cancer.sanger.ac.uk/cosmic/signatures, downloaded on January 15th, 2016). The following cut-offs were employed - AC1: 0; AC2: 0.03193533; AC3 0.1082812; AC4: 0.03266562; AC5: 0; AC6: 0.003351944; AC7: 0.0280924; AC8: 0.1814745; AC9: 0.09121354; AC10: 0.01686839; AC11: 0.07757047; AC12: 0.1937234; AC13: 0.01665886; AC14: 0.03158583; AC15: 0.03138468; AC16: 0.3521707; AC17: 0.004232865; AC18: 0.236441; AC19: 0.04058829; AC20: 0.04701714; AC21: 0.04009243; AC22: 0.03756267; AC23:

0.04005552; AC24: 0.03503163; AC25: 0.01658311; AC26: 0.0261603; AC27: 0.02200583; AC28: 0.03145322; AC29: 0.07188295; AC30: 0.03147694. This analysis was performed separately for hypermutated and non-hypermutated tumors. To identify signature enrichment and depletion patterns between shared and private, as well as between clonal and subclonal SNVs, a stratified signature analysis was performed using the function run_SMC() from YAPSA as described in Giessler et al, 2017. Briefly, the stratified analysis was performed as a multistep procedure: (1) a supervised analysis of mutational signatures was run without any stratification; (2) for every SNV in a sample, the stratum it belonged to was annotated; (3) for every stratum, a stratum-specific mutational catalog was built; and (4) a supervised NNLS (using lsei) with the constraint that the sum of exposures per stratum equals the exposures computed by the unstratified analysis was performed. Thereafter, enrichment and depletion patterns for all mutational signatures detected in step 1 were computed (Kruskal Wallis tests followed by pairwise posthoc Nemenyi tests corrected for multiple testing according to Benjamini and Hochberg (BH)) from the exposures in all strata with the help of the function stat_test_SMC().

Estimation of cell type composition

The enrichment of different cell types in tumor samples was evaluated from gene expression and DNA methylation data. Gene expression data were analyzed using the R package MCP-counter (Becht et al., 2016) on normalized read counts (logarithmic reads per kilobase per million). MCP-counter quantifies the abundance of eight immune cell populations, endothelial cells and fibroblasts from transcriptomic markers unique to either of the cell types. For each cell type a score is computed using the log2 geometric mean of the corresponding markers. Signatures of neural and blood cells in DNA methylation data were deconvolved using the function estimateCellCounts() implemented in the R packages minfi (Aryee et al., 2014). The function uses regression calibration to estimate the proportions of NeuN+ and NeuN- cells (FlowSorted.DLPFC.450k; Jaffe et al., 2018) and of different blood cells (FlowSorted.Blood.450k; Jaffe et al., 2018) from 450k/EPIC methylation data.

Driver gene selection

Putative driver mutations in glioblastoma were called based on the public database Intogen (Rubio-Perez et al., 2015; Gonzalez-Perez et al., 2013) and from recurrence. Among coding genes we searched for overrepresentation launching OncodriveFML (Mularoni et al., 2016) (with the coding regions file taken from https://bitbucket.org/bbglab/oncodrivefml/downloads/) and accepted genes at q-values ≤ 0.1 (Benjamini-Hochberg-corrected). We included non-coding genes if they were mutated in more than five patients. We in addition included *TERT* promoter mutations as likely drivers (Barthel et al., 2018), considering the region up to 2,500 base pairs upstream of the transcription start site.

Subclonality test

To determine subclonality in Figure 3A we modeled mutation counts as a sampling result from a binomial distribution. If sampling from a diploid locus, the sampling probability of a clonal mutation is 0.5x*CCF*, where *CCF* denotes the cancer cell fraction. We determined *CCF* during phylogenetic inference as described in detail below. Loci of different copy number states shift the sampling probability non-trivially. We excluded these loci, requiring a (normalized) coverage ratio between 0.9 and 1.1, and in addition excluded mutations on male sex chromosomes. Subclonality was then assumed if the measured mutation counts fell below the 95%-quantile of a binomial distribution with sampling probability 0.5x*CCF*, i.e.

$$P(X \leq r^{\mathrm{mut}}) = \sum_{k=0}^{r^{\mathrm{mut}}} \binom{r^{\mathrm{ref}} + r^{\mathrm{mut}}}{r^{\mathrm{mut}}} (0.5 \times CCF)^{r^{\mathrm{mut}}} \times (1 - 0.5 \times CCF)^{r^{\mathrm{ref}}} < 0.05 \quad (1)$$

where r^{ref} and r^{mut} denote the measured reference and mutated reads of a SNV or small insertion/deletion. In case of loss of heterozygosity, equation (1) is modified to

$$P(X \leq r^{\mathrm{mut}}) = \sum_{k=0}^{r^{\mathrm{mut}}} \binom{r^{\mathrm{ref}} + r^{\mathrm{mut}}}{r^{\mathrm{mut}}} (CCF)^{r^{\mathrm{mut}}} \times (1 - CCF)^{r^{\mathrm{ref}}} < 0.05. \quad (2)$$

We then used the classified mutations to estimate the number of subclonal mutations that appear clonal in a tumor sample due to incomplete tissue sampling. The fraction of these false positive clonal mutations can be assessed from mutations which were clonal in the primary and subclonal in the relapse sample. In fact, these mutations must have already been subclonal in the primary tumor as a transition from clonality to subclonality violates a single tumor origin. The false positive rate of clonal mutations due to incomplete tumor sampling thus reads

$$P(\text{subclonal in tumor} \mid \text{clonal in tumor sample}) = \frac{n_{\text{cPsR}}}{n_{\text{cPsR}} + n_{\text{cPcR}}}, \quad (3)$$

where n_{cPsR} is the number of mutations which were clonal in the primary and subclonal in the relapse sample and n_{cPcR} is the number of mutations which were clonal in both samples.

Phylogenetic inference

General considerations

Our phylogenetic analysis aims at inferring genetic subclones in pairs of primary and relapse tumors that are characterized by a combination of oncogenic drivers (encompassing mutations and copy number variations). Thus we will deduce robust features of the evolving phylogenetic tree of a tumor, including which driver mutations are in the common stem (i.e., occurred comparatively early) and which mutations are still subclonal at the times of surgical resection (and hence occurred later). Moreover, we will infer the common origins of subclones in the primary and relapse tumors, showing whether the relapse evolved from a unique subclone or multiple subclones in the primary tumor. For this analysis, we include all somatic variants specific to the tumor and do not restrict ourselves to known driver mutations. In particular, the majority of the 12,800 somatic SNVs recovered on average from a tumor are most likely neutral. However, the VAFs of neutral mutations in a tumor cell that harbors specific oncogenic drivers will grow in a coordinate manner as the subclone emerging from this cell expands, leading to many SNVs with the same VAF. In turn, this feature will facilitate the detection of the subclone in bulk WGS data.

The subclonal structure of a phylogenetic tree can be considered at different levels of resolution, from major branching down to single cells. Too fine a resolution is not of interest here because the vast majority of small subclones will be distinguished only by neutral somatic mutations that contain no information on tumor evolution by acquisition of driver mutations. Also, finely resolved phylogenetic trees cannot uniquely be reconstructed from bulk WGS data as mutations with small VAFs cannot be mapped uniquely to a large number of small subclones. These two considerations underscore our aim of reconstructing the occurrence of major drivers during genetic tumor evolution, which also motivates model selection (see below) favoring parsimonious trees that yield appropriate fits to the data. Importantly, the phylogenetic inference does not make assumptions on why two or more subclones coexist; coexistence could be a snapshot in the process of a newly evolved subclone outcompeting its predecessor(s) or be due to cooperation of subclones.

Assuming monoclonal origin of the tumor and neglecting the (very small) probability of backward evolution, a tumor can be visualized by a phylogenetic tree whose root corresponds to the founder cell of the tumor and whose tips represent the subclones in the tumor sample (Figure S3A). In accordance with the infinite sites hypothesis (Kimura, 1969), we assume that all SNVs and small insertions/deletions are singular events in the phylogenetic tree. Copy number changes often span larger regions so that the infinite sites model might not be adequate here, and a copy number change measured in both tumor samples could also be due to independent events. We address this by allowing up to two independent events changing copy number at a single locus, as described below in detail.

Mathematical description of the model

Assuming well-mixed samples, we model read counts in whole genome sequencing as a sampling result from a multinomial distribution, whose different categories represent the genetic subclones in the tumor. At each genomic locus the probability to sample reads from a distinct subclone scales with its relative size and its copy number state. Let $SC_i, i = \{1,...,K\}$ denote the *i*-th subclone in a

heterogeneous tumor with K subclones and $\mu = (\mu_1, ..., \mu_K)$ the proportion of cells in the sample originating from each subclone. Then, at each locus l, the probability to sample a read from SC_i in a Bernoulli trial can be written as

$$p_{i,l} = \mu_i \frac{\pi_{i,l}}{\sum_{i'=1}^{K} \mu_{i'} \pi_{i',l}} \tag{4}$$

where $\pi_{i,l} = (\pi_{1,l}, ..., \pi_{K,l})$ are integers that denote the copy numbers at locus l in each subclone. Now, let $s_i = (s_{i,l}, ..., s_{K,l} | s_{i,l} = \{0, ..., \pi_{i,l}\})$ be integers that denote the number of mutated alleles at locus l in each subclone. Accordingly, the probabilities to sample a read supporting the reference genome or a mutation, respectively, are given by

$$p_{i,l}^{\text{ref}} = p_{i,l}\left(1 - \frac{s_{i,l}}{\pi_{i,l}}\right) \quad \text{and} \quad p_{i,l}^{\text{mut}} = p_{i,l}\frac{s_{i,l}}{\pi_{i,l}}. \tag{5}$$

In bulk sequencing, the genomes from all subclones are intermingled. Consequently, r_L^{ref} reference reads and r_l^{mut} mutated reads at locus l originate from the K subclones in the sample, such that

$$r_l^{\text{ref}} = \sum_{i=1}^{K} r_{i,l}^{\text{ref}} \quad \text{and} \quad r_l^{\text{mut}} = \sum_{i=1}^{K} r_{i,l}^{\text{mut}}. \tag{6}$$

With this, the likelihood function for a measured number of reference and mutated reads at locus l is given by

$$\mathcal{L}_l(p_l \mid r_l^{\text{ref}}, r_l^{\text{mut}}) = P_l(r_l^{\text{ref}}, r_l^{\text{mut}} \mid p_l) = \frac{(r_l^{\text{ref}} + r_l^{\text{mut}})!}{\prod_{i=1}^{K} r_{i,l}^{\text{ref}}! r_{i,l}^{\text{mut}}!} \prod_{i=1}^{K} (p_{i,l}^{\text{ref}})^{r_{i,l}^{\text{ref}}} (p_{i,l}^{\text{mut}})^{r_{i,l}^{\text{mut}}}, \tag{7}$$

with the corresponding log-likelihood, l_l

$$l_l = \log \mathcal{L}_l = C + \sum_{i=1}^{K} \left[r_{i,l}^{\text{ref}} \log(p_{i,l}^{\text{ref}}) + r_{i,l}^{\text{mut}} \log(p_{i,l}^{\text{mut}}) \right], \tag{8}$$

where C is a constant that solely depends on the read counts.

Solution space and phylogenetic tree design

We assume that tumors evolve from a monoclonal origin and can be visualized by a phylogenetic tree. This restricts the solution space for subclonal inference as we will see in the following. Invoking the infinite sites hypothesis (Kimura, 1969), we require that the combinations of subclones carrying a mutation can be explained by a single event in the tree and are present on one of the two parental alleles only. We thus require

$$0 \leq s_l \leq \max(B_l, \pi_l - B_l), \tag{9}$$

where B_l is the number of B-alleles at locus l.

If an SNV collocates with a CNV, s_l becomes further restricted by the following criteria:

- if the copy number change precedes the mutation in the phylogenetic tree, the mutation can only be present on one allele
- if the mutation precedes the copy number change, the mutation must either be present on all A-alleles or on all B-alleles
- if the order of mutation and copy number change is unclear, the mutation can be present on any number of A- or B-alleles.

Model selection

To estimate μ, π and s, we first test a priori designed phylogenetic trees (Figure S3A). By treating normal tissue as an additional subclone, this automatically accounts for sample purity. Note that we designed the candidate trees in such a way that, in general, all subclones present in the primary sample are different from the ones in the relapse sample. However, this also comprises solutions in which the same subclone is present in both samples if the branches separating the two subclones are collapsed (Figure S3B). Similarly, these trees can also be collapsed into topologies of linear evolution (Figure S3B).

To select the most likely tree among the candidate topologies, we first require a good fit of the tumor stem. This is achieved by discarding trees in which more than 50% of the clonal (truncal) mutations are ambiguously mapped and, further, by discarding trees whose average squared error of clonal variant allele frequencies (VAFs) lay outside the 10% quantile of all candidate trees. We

then assess the likelihoods of the remaining trees with a modified Bayesian Information Criterion (BIC) (Chen and Chen, 2008). Briefly, the modified BIC incorporates increasing model complexity with increasing numbers of parameters:

$$BIC_\gamma = -2\log \mathcal{L}_n + v\log n + 2\gamma\log_\gamma, \quad (10)$$

where n is the number of data points, v the number of parameters and τ a parameter accounting for increasing model complexity weighted by Υ. In our case the number of data points is the number of readcounts ($r_l^{ref} + r_l^{mut}$), the number of parameters is the number of subclones, K, and τ is obtained by summing up all possible values of s. We choose $\Upsilon = 0.9$ to stringently incorporate the increasing model complexity when increasing the number of parameters.

We fit all 96 trees consisting of up to three tumor subclones per sample (Figure S3A) and identified the most likely solutions based on the reliability of the estimated clonal mutations and the modified *BIC*, as explained above.

Parameter estimation

Parameter estimations were performed on 500 mutated loci (including coding mutations and filled up by randomly chosen non-coding mutations). Subsequently, all mutations were mapped on the inferred tree structure. Upon fitting, copy number variations were manually inspected to identify large gains/losses and to adjust the solution in the case of inference problems due to high level amplifications or homozygous deletions.

We jointly estimated the parameters of matched primary and relapse samples with a maximum likelihood approach based on an expectation-maximization algorithm for different candidate trees. At a given evolutionary tree, a nested expectation step is followed by likelihood maximization from which the new input to the expectation step is generated. Both steps were iteratively repeated until convergence (required as $\sum(\mu^i - \mu^{i-1})^2 < 5\times10^{-4}$, where i is the index of the iteration). In order to identify the global maximum, optimization was repeated 100 times at

random starting conditions for each candidate tree. Expectation and maximization steps are described in detail in the following.

Expectation

We initiated the algorithm with randomly chosen values for μ. In all following iterations, the parameter estimates obtained in the maximization step were given as input. In each expectation step, the expected counts of mutated and reference reads per subclone were iteratively calculated at each mutated locus after inferring the copy number state as follows:

Expectation step: copy numbers

We assume that there is at most one dominating copy number change, $CN_{\text{aberr},l}$, per locus and sample. This change does not have to be present in all subclones, allowing for tumor heterogeneity. While we allow different copy number changes to dominate the primary and the relapse sample at a specific locus, we do not allow for multiple copy number changes per locus within a sample.

We determined $CN_{\text{aberr},l}$ from the normalized coverage ratios between tumor and blood along with the measured B-allele frequencies, BAF_l, in the tumor. To this end, we applied the following criteria:

- If no information on the coverage ratio was available, we assumed normal ploidy (2 on autosomes and female sex chromosomes, 1 on male sex chromosomes). If no information on the B-allele frequency was available, we assumed a B-allele frequency of 0.5 on autosomes and female sex chromosomes and of 0 on male sex chromosomes.
- We assumed that loci with coverage ratios in the interval [0.9, 1.1] and a *BAF* in the interval [0.45, 0.55] (or [0, 0.05] in case of male sex chromosomes) reflect normal copy number states, CN_{norm}, such that $CN_{\text{aberr},l} = CN_{\text{norm}} = 2$ on autosomes and female sex chromosomes and $CN_{\text{aberr},l} = CN_{\text{norm}} = 1$ on male sex chromosomes. The cutoffs were chosen based on the expected standard deviation of 8 % in Poisson distributed read counts at a coverage of 150x.

- At all other loci, we inferred the copy numbers and B-allele numbers by minimizing the squared errors between the expected and observed coverage ratios and B-allele frequencies. To this end, we started with a single copy number, which we iteratively increased. At each copy number we then tested different subclonal distributions $f \in F$ of the copy number change, where f is a vector $f = (f_1, f_2, \ldots, f_K)$, whose elements are binary indicators of a copy number change in the respective subclone, i.e. $f_i \in \{0,1\}$. The different combinations are restricted by the candidate tree and comprise solutions in which the same or two different copy number changes dominate the primary and the relapse sample, respectively. We computed the expected B-allele frequency for B-allele counts, $B_{\text{aberr},l}$, in the interval [0, $CN_{\text{aberr},l}$]

$$\mathrm{E}[\mathrm{BAF}_l] = \frac{\sum_i \left[f_i \mu_i B_{\text{aberr},l} + (1 - f_i) \mu_i B_{\text{norm}} \right]}{\sum_i \left[f_i \mu_i CN_{\text{aberr},l} + (1 - f_i) \mu_i CN_{\text{norm}} \right]}, \tag{11}$$

and chose the B-allele count that minimized the squared error between expected and observed B-allele frequencies. Likewise, we computed the expected coverage ratio, E[cr_l], with

$$\mathrm{E}[\mathrm{cr}_l] = \frac{\sum_i \left[f_i \mu_i CN_{\text{aberr},l} + (1 - f_i) \mu_i CN_{\text{norm}} \right]}{CN_{\text{norm}}}. \tag{12}$$

We aborted the algorithm once (E[BAF_l] - $BAF_{l,\text{obs}}$)^2 + (E[cr_l] - $cr_{l,\text{obs}}$)^2 < 0.01 (the threshold of 0.01 corresponds to the expected Poisson noise at sequencing depths of 150x) and determined $B_{i,l}$ and $\pi_{i,l}$ for each $i \in \{1,2,\ldots,\mathrm{K}\}$ to

$$\begin{aligned} B_{i,l} &= B_{\text{aberr},l} f_i + B_{\text{norm}} (1 - f_i) \\ \pi_{i,l} &= CN_{\text{aberr},l} f_i + CN_{\text{norm}} (1 - f_i) \end{aligned} \tag{13}$$

Note that while here only the most likely intratumoral distribution of a CNVs is selected, we accounted for alternative solutions for data representation and analysis as discussed below.

Expectation step: SNVs and small indels

Having inferred the copy number state at locus *l* we determined the expected read counts for each possible combination of s_l and π_l (which are predefined by the candidate tree) with

$$\mathrm{E}\left[r_{i,l}^{\mathrm{ref}}\right] = p'^{\mathrm{ref}}_{i,l}\, r_l^{\mathrm{ref}}, \qquad p'^{\mathrm{ref}}_{i,l} = \frac{p_{i,l}^{\mathrm{ref}}}{\sum_{i'=1}^{\mathrm{K}} p_{i',l}^{\mathrm{ref}}},$$
$$\mathrm{E}\left[r_{i,l}^{\mathrm{mut}}\right] = p'^{\mathrm{mut}}_{i,l}\, r_l^{\mathrm{mut}}, \qquad p'^{\mathrm{mut}}_{i,l} = \frac{p_{i,l}^{\mathrm{mut}}}{\sum_{i'=1}^{\mathrm{K}} p_{i',l}^{\mathrm{mut}}}, \tag{14}$$

where $p'_{i,l}{}^{\mathrm{ref}}$ and $p'_{i,l}{}^{\mathrm{mut}}$ are the conditional probabilities of a sampled reference or mutated read originating from SC_i provided that μ, s and π are known. We then computed the corresponding likelihood of μ, s and π as

$$\mathcal{L}_l(p_l \mid r_l^{\mathrm{ref}}, r_l^{\mathrm{mut}}) = \sum_{(r_{1,l}^{\mathrm{ref}}, r_{1,l}^{\mathrm{mut}})} \mathcal{L}_l(p_l \mid r_{i,l}^{\mathrm{ref}}, r_{i,l}^{\mathrm{mut}}) \tag{15}$$

and selected the solution with the highest likelihood (note that we considered alternative solutions for data representation and analysis also, if the best solution accounted for less than 90% of the total likelihood). Since DNA is fragmented before amplification and mapping, the read count distributions at different loci are independent of each other, so that the expectation step can be independently evaluated at each mutated locus. Of note, independence of measured coverage ratios is not guaranteed, since copy number variations can span multiple loci. This is already accounted for during segmentation and, thus, does not affect the inference procedure.

Maximization

We maximized the log-likelihood function (eqn. 8) at the expected readcount distribution (eqn. 14) w.r.t. μ. This was approached by summing up the log-likelihoods (eqn. 8) at each locus and by introducing the constraint $\sum_{i=1}^{\mathrm{K}} \mu_i = 1$ with a Lagrange multiplier before maximization:

$$\tilde{l} = \sum_l l_l + \lambda\left(1 - \sum_{i=1}^{K} \mu_i\right). \tag{16}$$

After inserting equation 5 into equation 16, deviation with respect to μ_i and λ yields:

$$\frac{\partial \tilde{l}}{\partial \mu_i} = \sum_l \left(\frac{1}{\mu_i}(r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}}) - \pi_{i,l} \frac{\sum_{i'=1}^{K} r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\sum_{i'=1}^{K} \mu_{i'} \pi_{i',l}} \right) - \lambda, \tag{17}$$

$$\frac{\partial \tilde{l}}{\partial \lambda} = 1 - \sum_{i=1}^{K} \mu_i. \tag{18}$$

We find the maximum of the log-likelihood by setting equations (17) and (18) equal to zero and solving for λ and μ:

$$\lambda \mu_i = \sum_l \left(r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}} - \pi_{i,l} \mu_i \frac{\sum_{i'=1}^{K} r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\sum_{i'=1}^{K} \mu_{i'} \pi_{i',l}} \right), \tag{19}$$

$$\sum_{i=1}^{K} \mu_i = 1. \tag{20}$$

Summing up eqn. 19 over all subclones yields

$$\lambda \sum_{i=1}^{K} \mu_i = \sum_{i=1}^{K} \sum_l \left(r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}} - \pi_{i,l} \mu_i \frac{\sum_{i'=1}^{K} r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\sum_{i'=1}^{K} \mu_{i'} \pi_{i',l}} \right). \tag{21}$$

With equation 20 this reduces to

$$\lambda = \sum_{i=1}^{K} \sum_l \left(r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}} - \pi_{i,l} \mu_i \frac{\sum_{i'=1}^{K} r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\sum_{i'=1}^{K} \mu_{i'} \pi_{i',l}} \right), \tag{22}$$

and consequently, by inserting equation 22 into equation 19, μ_i can be determined as

$$\mu_i = \frac{\sum_l r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}}}{\sum_l \left(\pi_{i,l} \frac{\sum_{i'=1}^{K} r_{i',l}^{\text{ref}} + r_{i',l}^{\text{mut}}}{\sum_{i'=1}^{K} \mu_{i'} \pi_{i',l}} \right) + \lambda}, \quad (23)$$

which reduces to

$$\mu_i = \frac{\sum_l \frac{r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}}}{\pi_{i,l}}}{\sum_{i=1}^{K} \sum_l \frac{r_{i,l}^{\text{ref}} + r_{i,l}^{\text{mut}}}{\pi_{i,l}}}. \quad (24)$$

Ambiguous solutions

To avoid bias in data interpretation due to ambiguous solutions, we accounted for three types of ambiguity:

- CNVs. If model inference suggested two independent copy number changes in primary and relapse tumor, respectively, but the squared error of a joined solution was less than twice the least squared error, we accounted for the joined solution in data analysis and interpretation. Likewise, if the squared error for a clonal copy number change was less than twice the error of a subclonal copy number change, we accounted for the clonal solution.
- SNVs. If the location of a mutation to the phylogenetic tree was non-unique, i.e., if the best solution carried less than 90% of the total likelihood at this locus, we sorted solutions by decreasing likelihood and accounted for all solutions that jointly yielded at least 90% of the total likelihood in data analysis and interpretation.
- Tree structure. We accounted for all solutions with $BIC_\gamma \leq \min(BIC_\gamma) + 10$.

Simulated data

To test the performance of phylogenetic tree reconstruction against a known "ground truth" computationally, we generated 100 test sets of up to three primary and recurrent subclones, respectively, according to the following algorithm:

- Sample between one and three primary and recurrent clones (K_{prim} and K_{rec})
- Sample the corresponding subclone sizes, μ_{prim} and μ_{rec}, at a minimal tumor cell content of 0.5
- Sample a random tree from Figure S3A
- Sample K_{prim} and K_{rec} nodes from this tree
- Sample between 1 and 200 mutations per subclone
- For each mutation sample a copy number change, π_l:
    - Normal copy numbers (two alleles) are weighted 15-fold
    - A unary copy number change (± 1) is weighted 2-fold
    - Copy number changes of up to +8 are weighted 1-fold each
- From the tree structure, sample the subclones that carry the copy number change, stored in the indicator vector π_l
- At each copy number change sample whether the mutation was affected by the copy number change; from this determine s_l
- Simulate the coverage ratios according to $cr_l = \frac{\sum_i \pi_{i,l} f_{i,l} \mu_i + 2(1 - f_{i,l})\mu_i}{2} + \mathcal{N}(\mu = 0.05, \sigma = 10^{-4})$ thus, adding Gaussian noise to the simulated coverage ratios
- Sample the read depth per locus from a Poisson distribution with $\lambda = 150$, corresponding to an average read depth of 150. Then, based on the read depth, sample the number of mutated reads from a binomial distribution with sampling probabilities according to equation (5).

Modeling of clonal dynamics

To link genetic evolution and tumor growth, we considered a deterministic model of tumor growth and mutation accumulation. We assumed exponential tumor growth at rate λ-δ,

where λ denotes the rate of cell division and δ the rate of cell death. The number of tumor cells, N, is modeled as

$$N(t) = e^{\lambda\left(1-\frac{\delta}{\lambda}\right)t}, \qquad N(0) = 1 \tag{25}$$

We implemented mutation accumulation at a constant rate μ per cell division, so that the number of mutations per cell, m, grows linearly with the number of divisions:

$$\dot{m} = \mu\lambda. \tag{26}$$

Accordingly, the time it takes to accumulate m mutations in a tumor cell is

$$T(m) = \frac{m}{\mu\lambda}. \tag{27}$$

Knowing μ, $m(t)$ and the number of tumor cells, $N(t)$, we can estimate the ratio between cell death and division without information on the tumor age. Rewriting equation 25 and inserting equation 27 yields:

$$\tilde{\delta} = \frac{\delta}{\lambda} = 1 - \frac{\mu}{m(t)} \log N(t), \tag{28}$$

where we defined $\tilde{\delta}$ as the death rate relative to cell divisions. To estimate $\tilde{\delta}$ from our data, μ, $m(t)$ and $N(t)$ were chosen as follows:

- We assumed a mutation rate between 0.26 x 10^{-8} to 1.06 x 10^{-8} mutations per cell division (Milholland et al., 2017). The upper boundary corresponds to a four-fold increase in somatic mutation rate, accounting for a drop in the proportion of the mutational signature 1 (linked to cell divisions) between tumor stem and tips (Figure S5K).
- We conservatively estimated the number of mutations per tumor cell at diagnosis as the average number of non-clonal mutations in all subclones of the primary tumor. Note that the mutations present in a subclone are shared by all cells of the subclone and thus have been accumulated in a single cell.
- Finally, we estimated a tumor size of approximately 10^9 - 10^{10} cells at diagnosis. This estimate has previously been suggested as the number of tumor cells per cubic centimeter (Del Monte, 2009; De Vita et al., 1975) and is in agreement with a brain tumor size of 20 - 80

1326 cm^3 and 10^{12} glia cells per brain (Goldberg-Zimring et al., 2005, Herculaono-Houzel et al.,
1327 2006, Milo et al., 2010, Pakkenberg et al., 1997) (corresponding to a brain size of
1328 approximately 1500 cm^3, see Drachman et al., 2005).
1329
1330 ***Selective advantage of TERT promoter mutations***
1331 To estimate the selective advantage provided by canonical *TERT* promoter mutations, we modeled
1332 its random acquisition and subsequent clonal expansion. The model samples the time of
1333 occurrence of the *TERT* promoter mutation from a Poisson process with given mutation rate. From
1334 this point onward, the *TERT* promoter-mutant subclone grows faster than *TERT* promoter-wildtype
1335 tumor cells. The model simulation is terminated when the tumor reaches the characteristic size of
1336 surgical resection. At this end point, the *TERT* promoter-mutant subclone may have reached
1337 clonality or may still be subclonal (although it would eventually become clonal if the tumor were
1338 allowed to grow for another year or so). Repeating this simulation will therefore result in a
1339 distribution for the *TERT* promoter-mutant tumor fraction that will be compared with the
1340 experimental data. The free parameter in this model is the selective advantage conferred by the
1341 *TERT* promoter mutation. This parameter will be adjusted such that the model matches the data,
1342 thus providing an estimate of the selective advantage conferred by the *TERT* promoter mutation.
1343 We began by assessing the probability of capturing the mutation after a certain number of
1344 cell divisions. The probability of at least one mutation (denoted as *TERT**) after n_{div} divisions is given
1345 by
1346 $$P(TERT^*) = 1-(1-2\mu_{\text{bp}})^{n_{\text{div}}}, \quad (29)$$
1347 where μ_{bp} denotes the per base substitution rate per cell division. The mutation probability
1348 is scaled by a factor of 2 accounting for the two canonical *TERT* promoter mutations recurrently
1349 observed in glioblastoma (chr5, 1,295,228 C>T and 1,295,250 C>T). The number of divisions, n_{div} is a
1350 time dependent function, modeled by
1351 $$\dot{n}_{\text{div}} = \lambda N, \quad (30)$$
1352 which, using equation 25 can be solved by

$$n_{\mathrm{div}} = \frac{1}{1-\tilde{\delta}}\left(e^{(1-\tilde{\delta})t} - 1\right). \quad (31)$$

Inserting this into equation 29 we obtain the probability distribution of at least one *TERT* promoter mutation at timepoint *t*. We modeled the effect of *TERT* promoter mutations by a decrease in cell death since these mutations have been linked to increased cellular survival (Chiba et al., 2017). Consequently, the expansion of the *TERT* mutated tumor fraction, N_{TERT}, is modeled by

$$N_{\mathrm{TERT}}(t) = \begin{cases} 0 & t < T_{0,\mathrm{TERT}} \\ e^{(1-\tilde{\delta}_{\mathrm{TERT}})(t-T_{0,\mathrm{TERT}})} & t \geq T_{0,\mathrm{TERT}} \end{cases} \quad (32)$$

The expansion of the tumor founder population is modeled by equation 25. The total tumor size consequently equals the sum of equations 25 and 32. Since *TERT* promoter mutations were found clonally in ≥ 2/3 of the tumors, the estimated tumor cell death fraction, $\tilde{\delta}$ (eqn. 28) must hold true for *TERT* mutated cells. Accordingly, we fixed the lower boundary of $\tilde{\delta}_{\mathrm{TERT}}$ based on equation 28 and for each $\min(\tilde{\delta}_{\mathrm{TERT}}) \leq \tilde{\delta}_{\mathrm{TERT}} \leq 0.999$ evaluated the death rate of the founder population, $\tilde{\delta}_0$. To do this we scanned the range $\tilde{\delta}_{\mathrm{TERT}} \leq \tilde{\delta}_0 \leq 0.999$ and at each value sampled 1,000 instances of the timepoint of the *TERT* promoter mutation, $T_{0,TERT}$, from equation 29 (assuming single base substitution rates of 0.26 x 10^{-8} to 1.06 x 10^{-8} (Milholland et al., 2017) and a tumor size of 10^9 cells, as before). We then evaluated the goodness of fit at each parameter combination by $\mathcal{X}^2$ minimization, comparing the mean and variance of the simulated and measured *TERT* mutated tumor fraction (the uncertainties of the mean and the variance in the measurement were estimated using bootstrapping with 10,000 resampling steps). Parameter combinations with $\mathcal{X}^2 \leq 5.99$, corresponding to the 95% confidence interval of the Chi²-distribution yield equally good fits and we selected the values of $\tilde{\delta}_0$ and $\tilde{\delta}_{\mathrm{TERT}}$ from $\min_{\tilde{\delta}_0, \tilde{\delta}_{\mathrm{TERT}}} | \mathcal{X}^2 \leq 5.99$

Finally, we assessed the selective advantage before and after the *TERT* promoter mutation, s_0 and s_{TERT}, as (Bozic et al., 2016)

$$s_0 = 1 - 2\frac{\tilde{\delta}_0}{1+\tilde{\delta}_0} \quad (33)$$

$$s_{TERT} = 1 - \frac{\tilde{\delta}_{TERT}}{1+\tilde{\delta}_{TERT}}\frac{1+\tilde{\delta}_0}{\tilde{\delta}_0} \quad (34)$$

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests

To measure differences in the mutational burden of primary and relapsed glioblastomas, we employed a two-sided Welch two sample t-test on the 17 non-hypermutated tumor pairs. The distribution of methylated and unmethylated *MGMT* promoters between tumors with mono- or oligoclonal relapse origins was assessed with a hypergeometric distribution (n=21, corresponding to the tumors of the discovery set). Differential gene expression between tumors and normal brain tissue was analyzed with the R package DESeq2 1.18.1 (Love et al., 2014); differential expression was accepted for Benjamini-Hochberg-corrected p values < 0.01.

All statistical parameters shown in figures are detailed in the figure legends.

Software

Phylogenetic inference and modeling of evolutionary dynamics were performed on R 3.3.1 (R Core Team, 2017), using the libraries doParallel, foreach and phangorn. The R libraries AnnotationsDbi, DESeq2, FlowSorted.DLPFC.450k, FlowSorted.Blood.450k, GenomicRanges, Hmisc, Homo.sapiens, MCP-counter, minfi, phangorn, phyloTop, phytools, YAPSA and xlsx were used for data analysis and the libraries beeswarm, colorlovers, colorspace, ComplexHeatmaps, ggplot2, igraph, pheatmap, plotrix and RColorBrewer were used for plotting. Plots were redesigned in Adobe Illustrator CS5. All other software used is mentioned in METHOD DETAILS.

DATA AND SOFTWARE AVAILABILITY

Sequencing data have been deposited in the European Genome-phenome Archive (EGA) under the accession number EGAS00001003184. R codes for phylogenetic inference and modeling of clonal dynamics are available as supplementary files to this paper.

Supplementary Tables and Figure legends

Table S1. Clinical data of the investigated glioblastoma patient cohort.

Figure S1. Tumor and immune cell content, related to Figure 1.

(A) Estimated tumor cell content in primary and recurrent samples from the discovery set assessed with phylogenetic inference.

(B) Immune cell composition from gene expression profiles (samples are ordered pairwise: left, primary; right, relapse tumor).

(C) Estimated cell type composition of NeuN-positive and –negative cells (left) and immune cells (right) from methylation patterns (boxes specify interquartile ranges; horizontal lines indicate medians; whiskers extend to the most extreme datapoints that are no more than 1.5 times the interquartile range; outliers are represented by circles; data is from 98 samples from both the discovery and the validation set).

(D) Methylation subtypes in primary and recurrent tumors (from 47 pairs of both the discovery and the validation set for which methylation subtypes of both samples could be determined).

Figure S2. Mutational profile of primary and relapsed tumors, related to Figure 2.

(A) Potentially druggable mutations in primary and relapsed tumors of the discovery set (excluding hypermutated cases). Shown are mutations (non-synonymous/stop-gain SNVs, homozygous deletions, small insertions/deletions and structural variants) found in at least one tumor. Vertical

lines separate tumor pairs (left primary, right relapsed); methylation subtypes are indicated at the bottom.

(B) Mutation distribution in the discovery and the validation set for four prominent oncogenes (*TERT* promoter, *PTEN*, *EGFR*, *TP53*; excluding samples analyzed with both methods).

Figure S3. Phylogenetic inference from whole genome sequencing data, related to Figure 3 and STAR Methods.

(A-B) Candidate trees for phylogenetic inference.

(A) Tree templates used to infer the phylogenetic structure. Three examples of unique distributions of primary and relapse samples are shown for each tree, if assuming maximally three subclones per sample.

(B) Trees can be collapsed to obtain solutions, in which primary and relapse samples share clones, or to topologies representing linear evolution.

(C-L) Phylogenetic inference on simulated data. Shown are the results from 100 simulations of up to three subclones per primary and relapse sample, respectively.

(C) True and inferred tumor cell content with Pearson's correlation coefficient.

(D) True and inferred clone size with Pearson's correlation coefficient. Shown are only cases, in which the correct number of subclones had been called.

(E) Deviation from the true number of subclones.

(F, G) False negative and false positive clonal mutations, if neglecting subclones ≤10%.

(H) True and inferred difference between the most recent common ancestors of primary and relapse sample, measured in mutation counts. Pearson's correlation coefficient is indicated in the plot.

(I) True and inferred mean number of mutations per subclone since the most recent common ancestor population of the tumor.

(K) Sensitivity of detecting a subclone in dependence of the relative subclone size.

(L) Specificity of detecting a subclone in dependence of the relative subclone size.

(M) Comparison of tumor cell content estimates from phylogenetic inference and from ACEseq. Pearson's correlation coefficient is indicated in the plot.

Figure S4. Phylogenetic trees of the discovery set, related to Figure 3.

Trees are designed as explained in Figure 3 (i.e. vertical branch lengths scale with the number of mutations, circle sizes scale with relative subclone sizes; mutations in driver genes and copy number changes of chromosomes 7, 9 and 10 are indicated; yellow circles, primary tumor; orange circles, recurrent tumor).

(A-N) Cases with an oligoclonal origin of the relapsed tumor. Panels (K-N) correspond to hypermutated cases, where only clonal mutations are indicated.

(O-S) Cases with a more monoclonal origin of the relapsed tumor. The tree shown in (Q) was due to low tumor cell content selected based on the purity estimate by ACEseq.

Shown are all tumors of the discovery set excluding the three tumors shown in Fig. 3.

Figure S5. Subclonality of *TERT* promoter mutations and evolutionary dynamics during tumor growth, related to Figures 4 and 5.

(A) Variant allele frequencies (VAFs) of canonical *TERT* promoter mutations (black), germline mutations (grey) and expected VAFs if sampling from a binomial distribution (red). VAFs of *TERT* promoter mutations were adjusted with the average estimate of tumor cell content from phylogenetic inference and ACEseq (as shown in Fig. S3M; tumors with non-neutral copy number at the *TERT* promoter were excluded, yielding data from 16 tumor pairs). Germline mutations were chosen from all tumors of the discovery set at loci with read coverage within the inter-quartile-range of the coverage at the *TERT* promoter (IQR=[106,140], assessed from the 16 pairs of which the VAFs at the *TERT* promoter mutation are shown) and down-sampled to 1,000 loci for better visualization. Expected germline VAFs (red) were computed from the average of binomial distributions if drawing samples of the measured read coverages with success probability 0.5.

(B) Kernel density estimates of the VAFs at heterozygous germline mutations (grey) and *TERT* promoter mutations (black; the solid line shows the average, the opaque area the 95% confidence interval of the kernel density estimate after sampling the tumor cell contents 10,000 times from the estimates from phylogenetic inference and ACEseq; displayed data are downsampled as in (A)).
(C) Reference sequence at the *TERT* promoter (red, canonical SNVs; black, germline SNP).
(D,E) Measured (points) and predicted (lines) cumulative distributions of VAFs at canonical *TERT* promoter mutations and a germline SNP (solid lines, binomial distribution at the lower 25% quantile of the measured coverage; red lines, 95% lower confidence bound at the lower 25% quantile of the measured coverage).
(D) Heterozygous germline SNP (1295349 A>G) in the discovery set (median coverage=140x, IQR=[130x,156x]; displayed data are from eight tumor pairs in which the polymorphism was found).
(E) Canonical *TERT* promoter mutations in the discovery set (VAFs are corrected for tumor cell content; median coverage=97x, IQR=[77x,114x] after correcting for tumor cell content; tumors with non-neutral copy number at the *TERT* promoter were excluded, yielding data from 16 tumor pairs).
(F) Immunohistochemical staining for ATRX in a primary and relapsed tumor of a selected patient with a frame-shift *ATRX* mutation and a non-canonical *TERT* promoter mutation in both tumors (antibody binding visualized with diaminobenzidine as chromogen (brown), all sections counterstained with hematoxylin (blue), original microscopical magnification: 400x; in total samples from four patients were stained of which two are exemplarily shown in (F) and (H)). Note loss of nuclear ATRX expression in both tumor samples. Interspersed vascular endothelia serve as endogenous control with retained nuclear ATRX expression.
(G) *TERT* expression (RNA-seq read counts normalized by library size) in normal brain tissue samples (grey dots; normal brain tissue data were generated by The Cancer Genome Atlas (TCGA) research network: http://cancergenome.nih.gov/), a tumor pair with non-canonical *TERT* promoter mutation (red dots) and in tumors with canonical *TERT* promoter mutations (black dots). ATRX expression of the tumor pair with non-canonical *TERT* promoter mutation is shown in (F).

(H) Nuclear ATRX expression in a primary tumor without *ATRX* mutation and retention of nuclear ATRX expression in the relapsed tumor of the same patient with a non-synonymous SNV in *ATRX*. Staining and magnification as in (F).

(G) *TERT* expression (RNA-seq read counts normalized by library size) in normal brain tissue samples (grey dots; normal brain tissue data were generated by The Cancer Genome Atlas (TCGA) research network: http://cancergenome.nih.gov/), a tumor pair with non-canonical *TERT* promoter mutation (red dots) and in tumors with canonical *TERT* promoter mutations (black dots). ATRX expression of the tumor pair with non-canonical *TERT* promoter mutation is shown in (C).

(I) Mean number of mutations per primary subclone accumulated after the most recent common ancestor of the tumor (shown are the tumors of the discovery set, excluding hypermutated samples; red line, median).

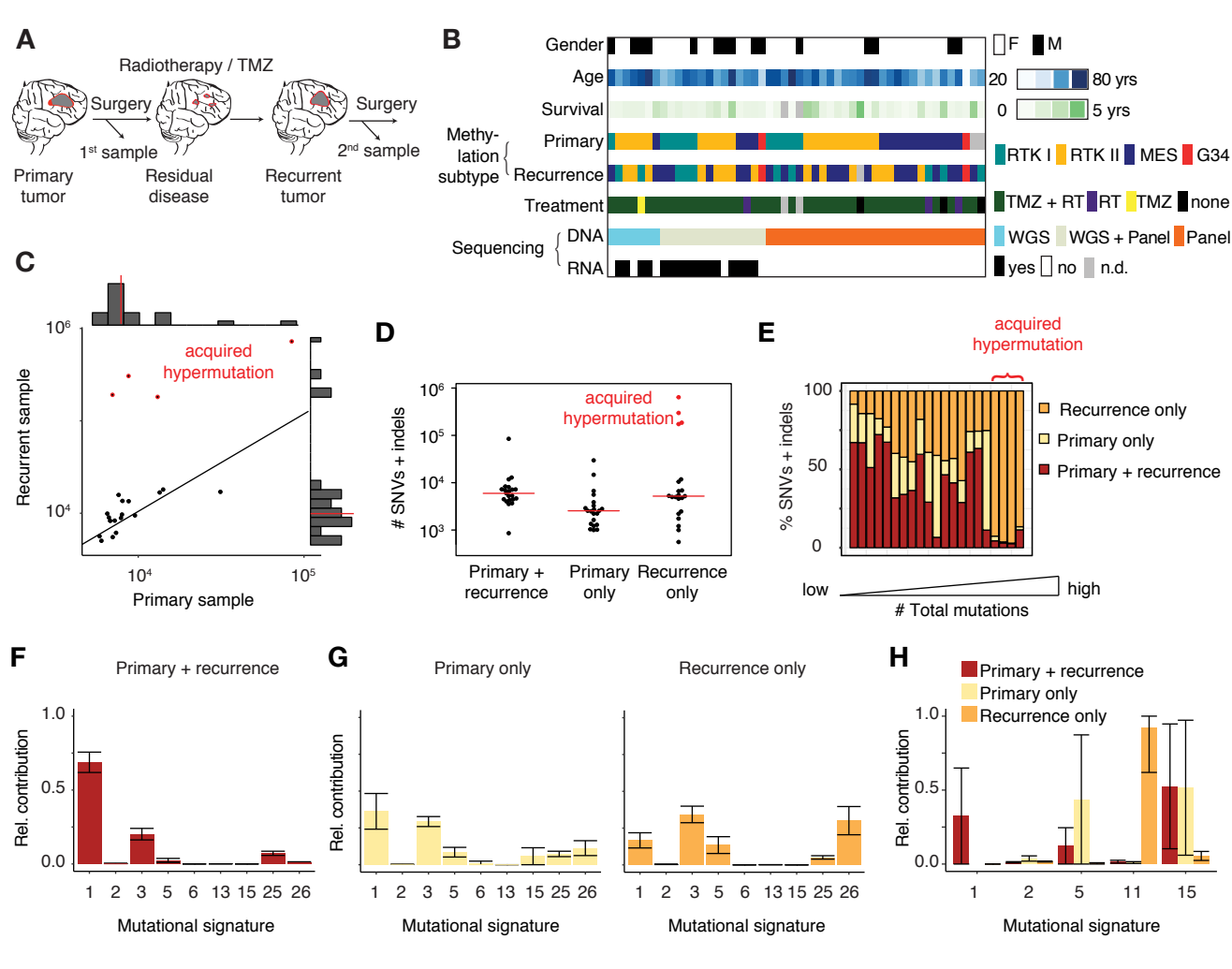
(K) Relative contribution of the clock-like mutational signature AC1 to mutations residing at the trunk and the tips of the phylogenetic trees (separately for primary and relapse subclones from cases of the discovery set, excluding hypermutated samples; red lines, median).

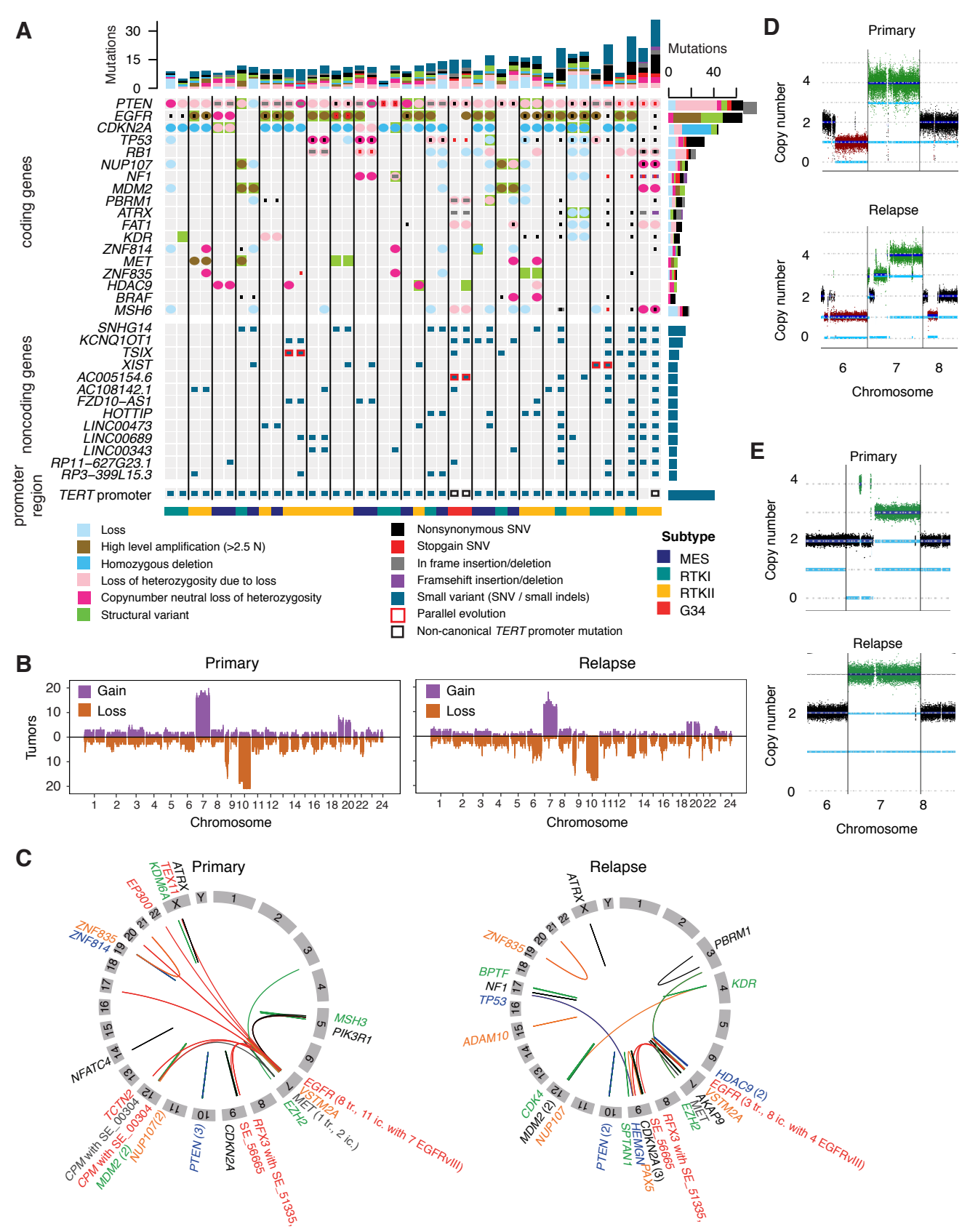
(L) Fraction of cell death per cell division required to reconcile an observed number of subclonal mutations with a tumor size of 10^9 cells (red line, median number of mutations per subclone observed in the data if excluding hypermutated cases; shaded area, inter quartile range).

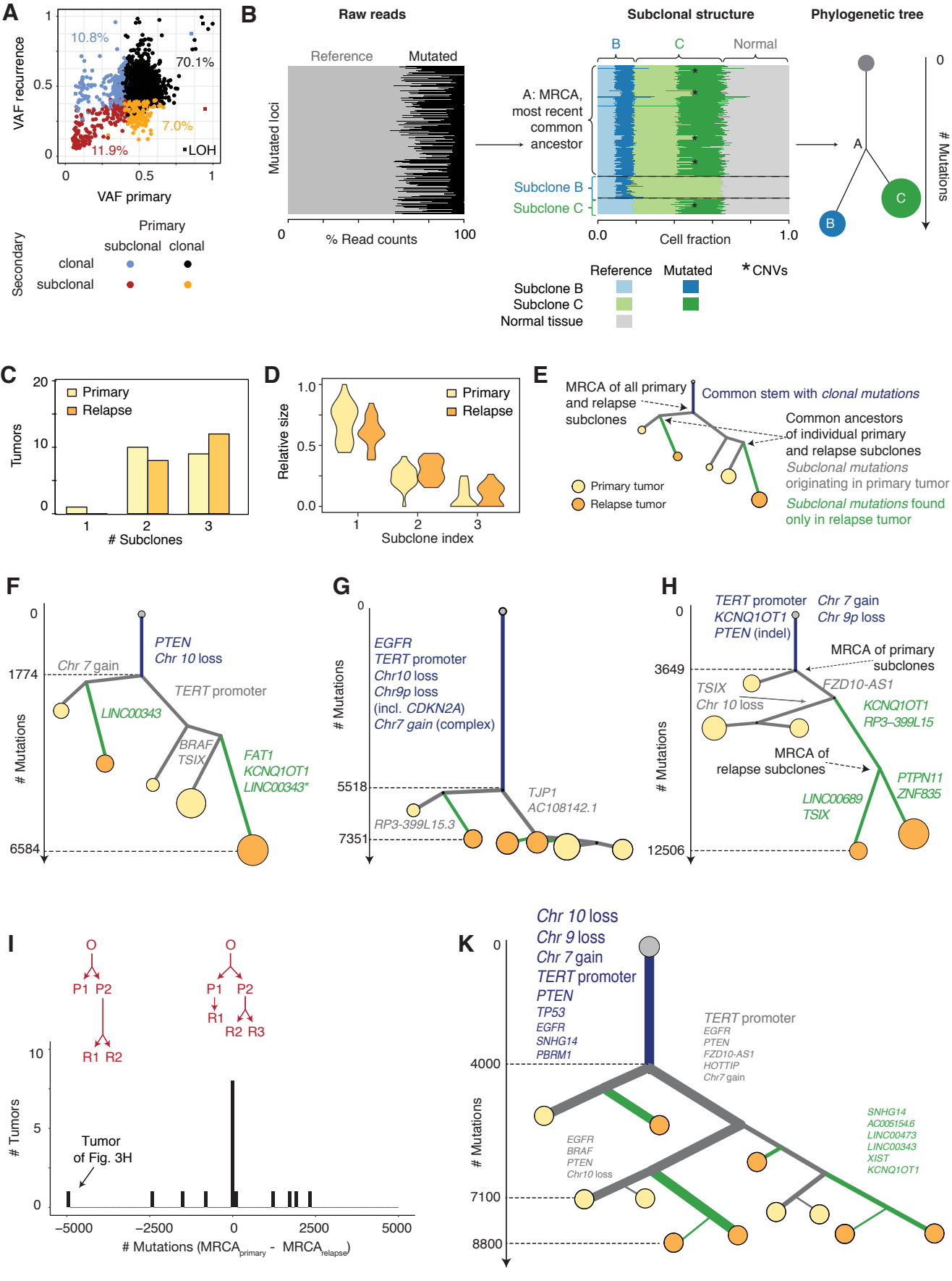
(M) Genes associated with positive regulation of cell death (GSEA: GO:0010942, from http://software.broadinstitute.org/gsea/msigdb/), which were differentially expressed ($p<0.01$ after Benjamini-Hochberg correction) between tumors (n=36) and normal brain tissue controls (n=5, normal brain tissue data were generated by The Cancer Genome Atlas (TCGA) research network: http://cancergenome.nih.gov/). Shown are log2 fold changes between tumor and control.

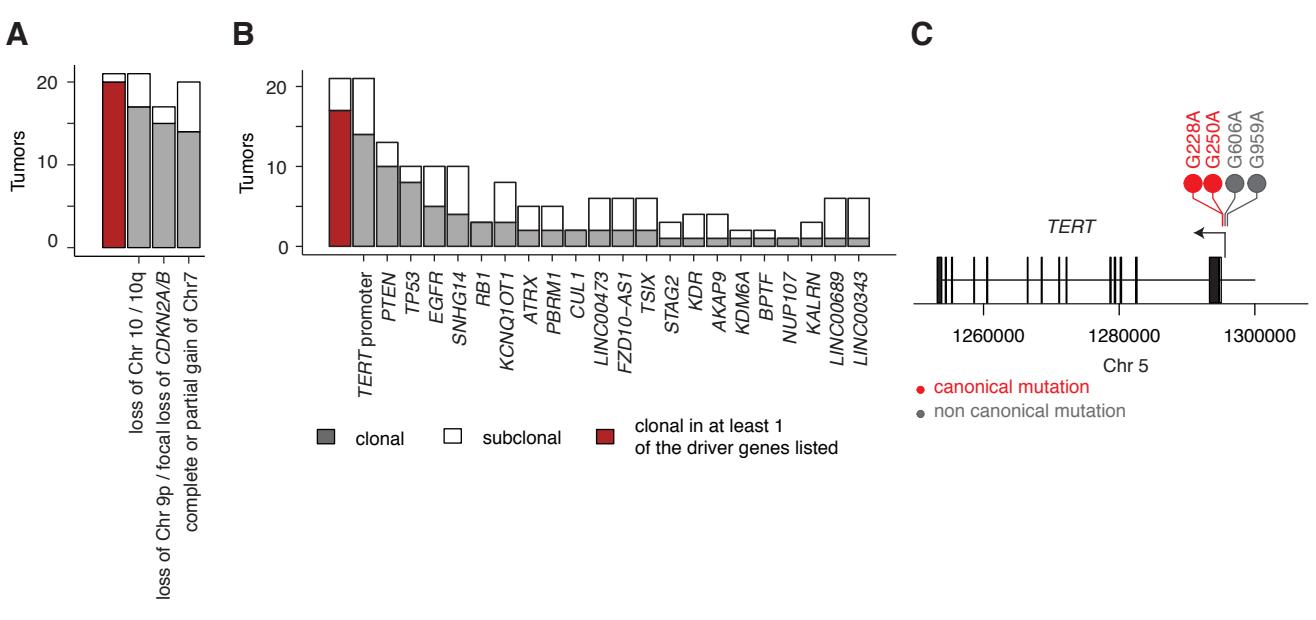
(N, O) Residual sum of squares when comparing the mean and variance of measured and modeled *TERT* promoter-mutant tumor fraction at different values of δ_0 and δ_{TERT} (grey dashed lines, bisetrices; data are from 19 tumors of the discovery set with canonical *TERT* promoter mutation). For each point the simulated mean and variance of 1000 simulations were compared to the measurement. Shown are the results if implementing *TERT* promoter mutations at the somatic

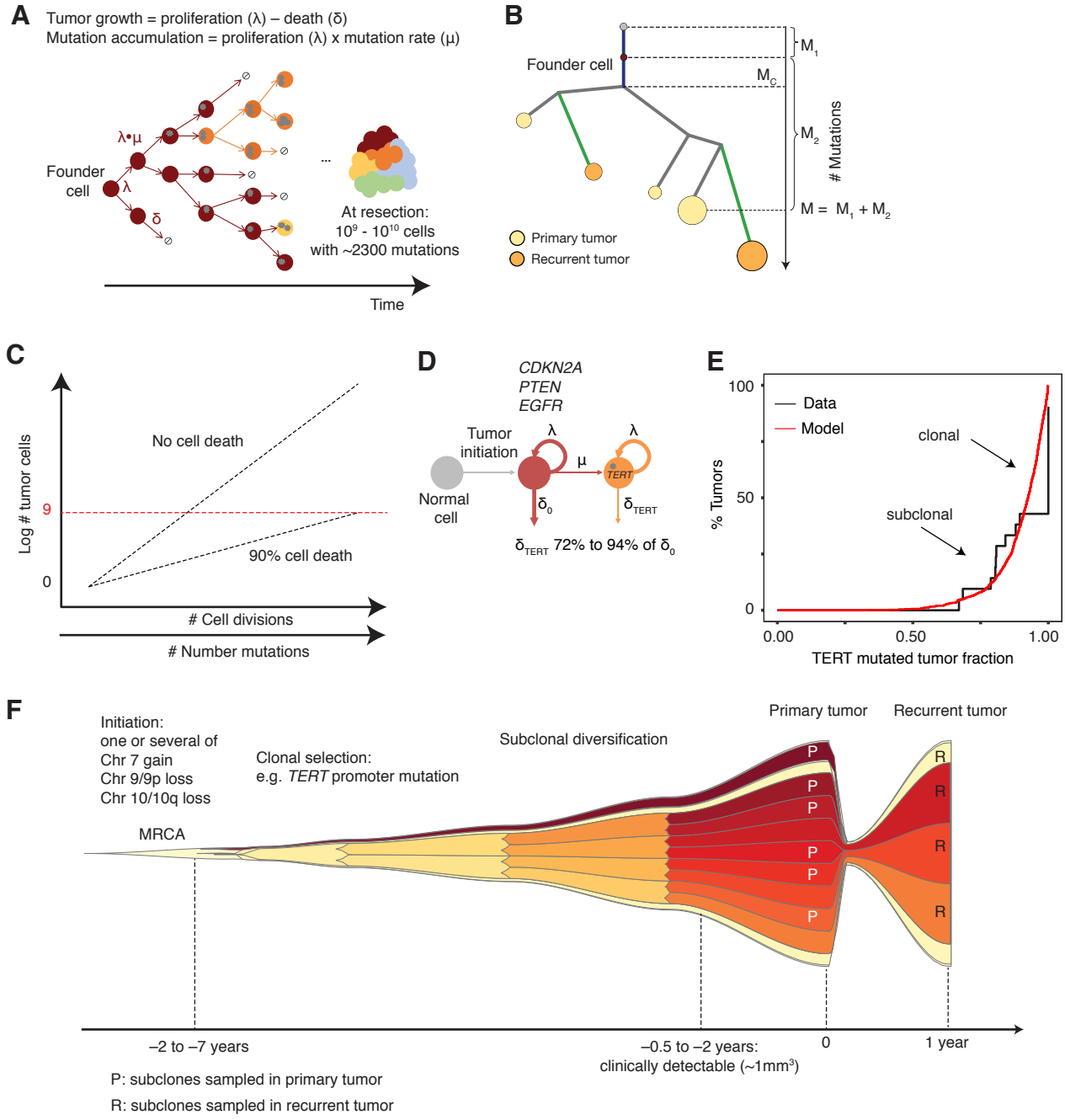
mutation rate (N) and four-fold the somatic mutation rate (O). Estimates lying within the 95% confidence interval (corresponding to a $\chi^2 \leq 5.99$) are shaded in dark red. These parameter combinations give equally good fits. Black lines mark the fractions of cellular death required to explain tumor sizes of 10^9 cells (as shown in (L)). This fraction is required for both the *TERT* promoter-mutant tumor fraction and the tumor fraction without *TERT* promoter mutation. Since $\delta_{TERT} \leq \delta_0$, it is sufficient to require the cutoff for δ_{TERT}, corresponding to the horizontal line. Thus parameter estimates explaining both, the *TERT* promoter-mutant tumor fraction and mutation accumulation in tumors of realistic sizes, intersect the horizontal black line and are colored in dark red. The corresponding values for δ_0 are marked by vertical lines.


A
Radiotherapy / TMZ
Surgery
Surgery
1st sample
2nd sample
Primary tumor
Residual disease
Recurrent tumor
B
Gender
Age
Survival
Methy-lation subtype
Primary
Recurrence
Treatment
Sequencing
DNA
RNA
F
M
20
80 yrs
0
5 yrs
RTK I
RTK II
MES
G34
TMZ + RT
RT
TMZ
none
WGS
WGS + Panel
Panel
yes
no
n.d.
C
acquired hypermutation
Recurrent sample
Primary sample
D
acquired hypermutation
SNVs + indels
Primary + recurrence
Primary only
Recurrence only
E
acquired hypermutation
% SNVs + indels
Recurrence only
Primary only
Primary + recurrence
low
high
Total mutations
F
Primary + recurrence
Rel. contribution
Mutational signature
G
Primary only
Recurrence only
Rel. contribution
Mutational signature
H
Primary + recurrence
Primary only
Recurrence only
Rel. contribution
Mutational signature


A
Mutations
30
15
0
Mutations
0
40
coding genes
noncoding genes
promoter region
PTEN
EGFR
CDKN2A
TP53
RB1
NUP107
NF1
MDM2
PBRM1
ATRX
FAT1
KDR
ZNF814
MET
ZNF835
HDAC9
BRAF
MSH6
SNHG14
KCNQ1OT1
TSIX
XIST
AC005154.6
AC108142.1
FZD10-AS1
HOTTIP
LINC00473
LINC00689
LINC00343
RP11-627G23.1
RP3-399L15.3
TERT promoter
Loss
High level amplification (>2.5 N)
Homozygous deletion
Loss of heterozygosity due to loss
Copynumber neutral loss of heterozygosity
Structural variant
Nonsynonymous SNV
Stopgain SNV
In frame insertion/deletion
Framshift insertion/deletion
Small variant (SNV / small indels)
Parallel evolution
Non-canonical TERT promoter mutation
Subtype
MES
RTKI
RTKII
G34
D
Primary
Relapse
Copy number
Chromosome
E
Primary
Relapse
Copy number
Chromosome
B
Primary
Relapse
Gain
Loss
Tumors
Chromosome
C
Primary
Relapse
ATRX
KDM6A
TEX11
EP300
ZNF835
ZNF814
NFATC4
TCTN2
CPM with SE_00304
CPM with SE_00304
MDM2 (2)
NUP107(2)
PTEN (3)
CDKN2A
SE_56665
RFX3 with SE_51335,
MSH3
PIK3R1
EGFR (8 tr., 11 ic. with 7 EGFRvIII)
TSTM2A
MET (1 tr., 2 ic.)
EZH2
ATRX
ZNF835
BPTF
NF1
TP53
ADAM10
CDK4
MDM2 (2)
NUP107
PTEN (2)
SE_56665
RFX3 with SE_51335,
CDKN2A (3)
HEMGN
SPTAN1
PAX5
MET
AKAP9
EZH2
TSTM2A
EGFR (3 tr., 8 ic. with 4 EGFRvIII)
HDAC9 (2)
KDR
PBRM1


A
VAF recurrence
VAF primary
10.8%
70.1%
7.0%
11.9%
LOH
Primary
subclonal clonal
Secondary
clonal
subclonal
B
Raw reads
Reference
Mutated
Mutated loci
% Read counts
Subclonal structure
B
C
Normal
A: MRCA, most recent common ancestor
Subclone B
Subclone C
Cell fraction
Phylogenetic tree
A
B
C
Mutations
Reference
Mutated
*CNVs
Subclone B
Subclone C
Normal tissue
C
Tumors
Primary
Relapse
Subclones
D
Relative size
Primary
Relapse
Subclone index
E
MRCA of all primary and relapse subclones
Common stem with clonal mutations
Common ancestors of individual primary and relapse subclones
Subclonal mutations originating in primary tumor
Subclonal mutations found only in relapse tumor
Primary tumor
Relapse tumor
F
Mutations
PTEN
Chr 10 loss
Chr 7 gain
TERT promoter
LINC00343
BRAF
TSIX
FAT1
KCNQ1OT1
LINC00343*
1774
6584
G
Mutations
EGFR
TERT promoter
Chr10 loss
Chr9p loss
(incl. CDKN2A)
Chr7 gain (complex)
TJP1
AC108142.1
RP3-399L15.3
5518
7351
H
Mutations
TERT promoter
KCNQ1OT1
PTEN (indel)
Chr 7 gain
Chr 9p loss
MRCA of primary subclones
TSIX
Chr 10 loss
FZD10-AS1
KCNQ1OT1
RP3–399L15
MRCA of relapse subclones
PTPN11
ZNF835
LINC00689
TSIX
3649
12506
I
O
P1 P2
R1 R2
O
P1 P2
R1
R2 R3
Tumors
Tumor of Fig. 3H
Mutations (MRCA primary - MRCA relapse)
K
Mutations
Chr 10 loss
Chr 9 loss
Chr 7 gain
TERT promoter
PTEN
TP53
EGFR
SNHG14
PBRM1
TERT promoter
EGFR
PTEN
FZD10-AS1
HOTTIP
Chr7 gain
SNHG14
AC005154.6
LINC00473
LINC00343
XIST
KCNQ1OT1
EGFR
BRAF
PTEN
Chr10 loss
4000
7100
8800


A
Tumors
20
10
0
loss of Chr 10 / 10q
loss of Chr 9p / focal loss of CDKN2A/B
complete or partial gain of Chr7
B
Tumors
20
10
0
TERT promoter
PTEN
TP53
EGFR
SNHG14
RB1
KCNQ1OT1
ATRX
PBRM1
CUL1
LINC00473
FZD10-AS1
TSIX
STAG2
KDR
AKAP9
KDM6A
BPTF
NUP107
KALRN
LINC00689
LINC00343
clonal
subclonal
clonal in at least 1
of the driver genes listed
C
G228A
G250A
G606A
G959A
TERT
1260000
1280000
1300000
Chr 5
canonical mutation
non canonical mutation


A
Tumor growth = proliferation (λ) – death (δ)
Mutation accumulation = proliferation (λ) x mutation rate (μ)
Founder cell
λ•μ
λ
δ
At resection:
10^9 - 10^10 cells
with ~2300 mutations
Time
B
Founder cell
M1
MC
M2
# Mutations
M = M1 + M2
Primary tumor
Recurrent tumor
C
Log # tumor cells
No cell death
9
0
90% cell death
Cell divisions
Number mutations
D
CDKN2A
PTEN
EGFR
Tumor initiation
λ
λ
μ
TERT
Normal cell
δ0
δTERT
δTERT 72% to 94% of δ0
E
100
50
0
% Tumors
Data
Model
clonal
subclonal
0.00
0.50
1.00
TERT mutated tumor fraction
F
Initiation:
one or several of
Chr 7 gain
Chr 9/9p loss
Chr 10/10q loss
Clonal selection:
e.g. TERT promoter mutation
Subclonal diversification
Primary tumor
Recurrent tumor
MRCA
P
R
–2 to –7 years
–0.5 to –2 years:
clinically detectable (~1mm3)
0
1 year
P: subclones sampled in primary tumor
R: subclones sampled in recurrent tumor

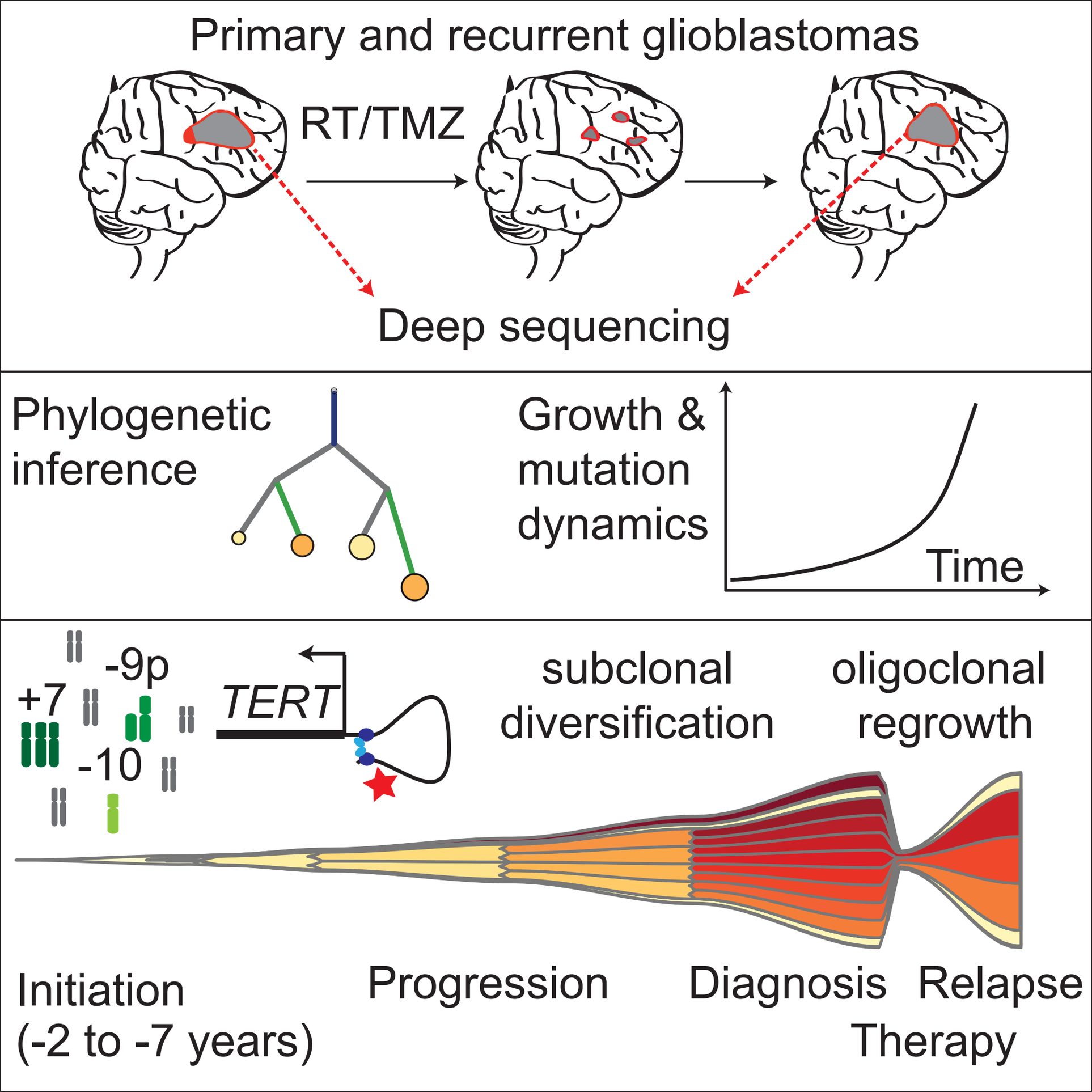
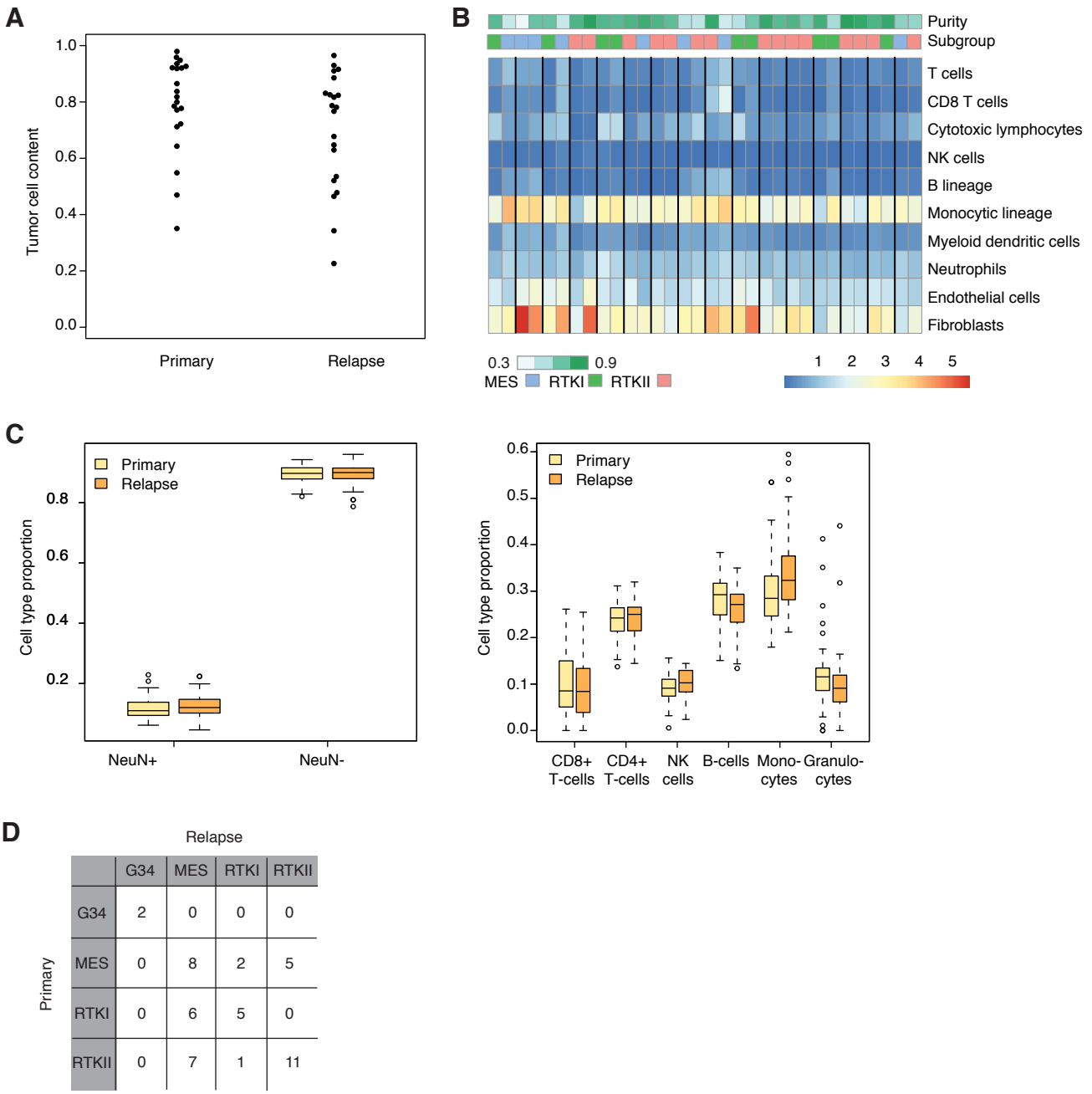

Primary and recurrent glioblastomas
RT/TMZ
Deep sequencing
Phylogenetic inference
Growth & mutation dynamics
Time
-9p
+7
-10
TERT
subclonal diversification
oligoclonal regrowth
Initiation (-2 to -7 years)
Progression
Diagnosis
Relapse
Therapy

Table S1. Clinical data of the investigated glioblastoma patient cohort.

	Entire cohort (n=50)	WGS cohort (n=21)
Age at diagnosis (years)		
Median (range)	60 (14 – 80)	60 (33 – 78)
Gender		
Male	35 (70%)	11 (52%)
Female	15 (30%)	10 (48%)
Tumor location		
Frontal	11 (22%)	4 (19%)
Temporal	18 (36%)	6 (29%)
Parietal	5 (10%)	3 (14%)
Occipital	3 (6%)	1 (5%)
More than 1 cerebral lobe	13 (26%)	7 (33%)
Local relapse	48 (96%)	20 (95%)
Extent of initial surgery		
Gross total resection	20 (40%)	11 (52%)
Subtotal resection	29 (48%)	10 (48%)
No data	1 (2%)	-
***MGMT* promoter methylation, primary tumor**		
Methylated	21(42%)	11 (52%)
Unmethylated	27 (54%)	10 (48%)
No data	2 (4%)	-
***MGMT* promoter methylation, recurrent tumor**		
Methylated	23 (46%)	12 (57%)
Unmethylated	27 (54%)	9 (43%)
No data	-	-
DNA methylation subgroup, primary tumor		
RTK I	11 (22%)	6 (29%)
RTK II	19 (38%)	10 (48%)
Mesenchymal	16 (32%)	4 (19%)
H3-G34	2 (4%)	1 (5%)
No data	2 (4%)	-
DNA methylation subgroup, recurrent tumor		
RTK I	9 (18%)	4 (19%)
RTK II	16 (32%)	9 (43%)
Mesenchymal	22 (44%)	7 (33%)
H3-G34	2 (4%)	1 (5%)
No data	1 (2%)	-
First-line therapy		
Radiotherapy alone	5 (10%)	1(5%)
Temozolomide alone	1 (5%)	1 (5%)
Radiotherapy plus temozolomide	42 (84%)	19 (90%)
No therapy beyond surgery	3 (6%)	-
Survival data (days), median (range)		
Interval between first and second surgery	290 (35 -1500)	280 (46 - 994)
Overall survival	518 (147 – 3168)	580 (261 - 1783)
Patients alive at last follow-up	25 (50%)	10 (48%)

Abbreviations used: WGS, whole genome sequencing; RTK I, receptor tyrosine kinase I group; RTK II, receptor tyrosine kinase II group; H3-G34, H3F3A-G34-mutant group.



D

Primary \ Relapse	G34	MES	RTKI	RTKII
G34	2	0	0	0
MES	0	8	2	5
RTKI	0	6	5	0
RTKII	0	7	1	11

Figure S1. Tumor and immune cell content, related to Figure 1.
(A) Estimated tumor cell content in primary and recurrent samples from the discovery set assessed with phylogenetic inference.
(B) Immune cell composition from gene expression profiles (samples are ordered pairwise: left, primary; right, relapse tumor).
(C) Estimated cell type composition of NeuN-positive and –negative cells (left) and immune cells (right) from methylation patterns (boxes specify interquartile ranges; horizontal lines indicate medians; whiskers extend to the most extreme datapoints that are no more than 1.5 times the interquartile range; outliers are represented by circles; data is from 98 samples from both the discovery and the validation set).
(D) Methylation subtypes in primary and recurrent tumors (from 47 pairs of both the discovery and the validation set for which methylation subtypes of both samples could be determined).

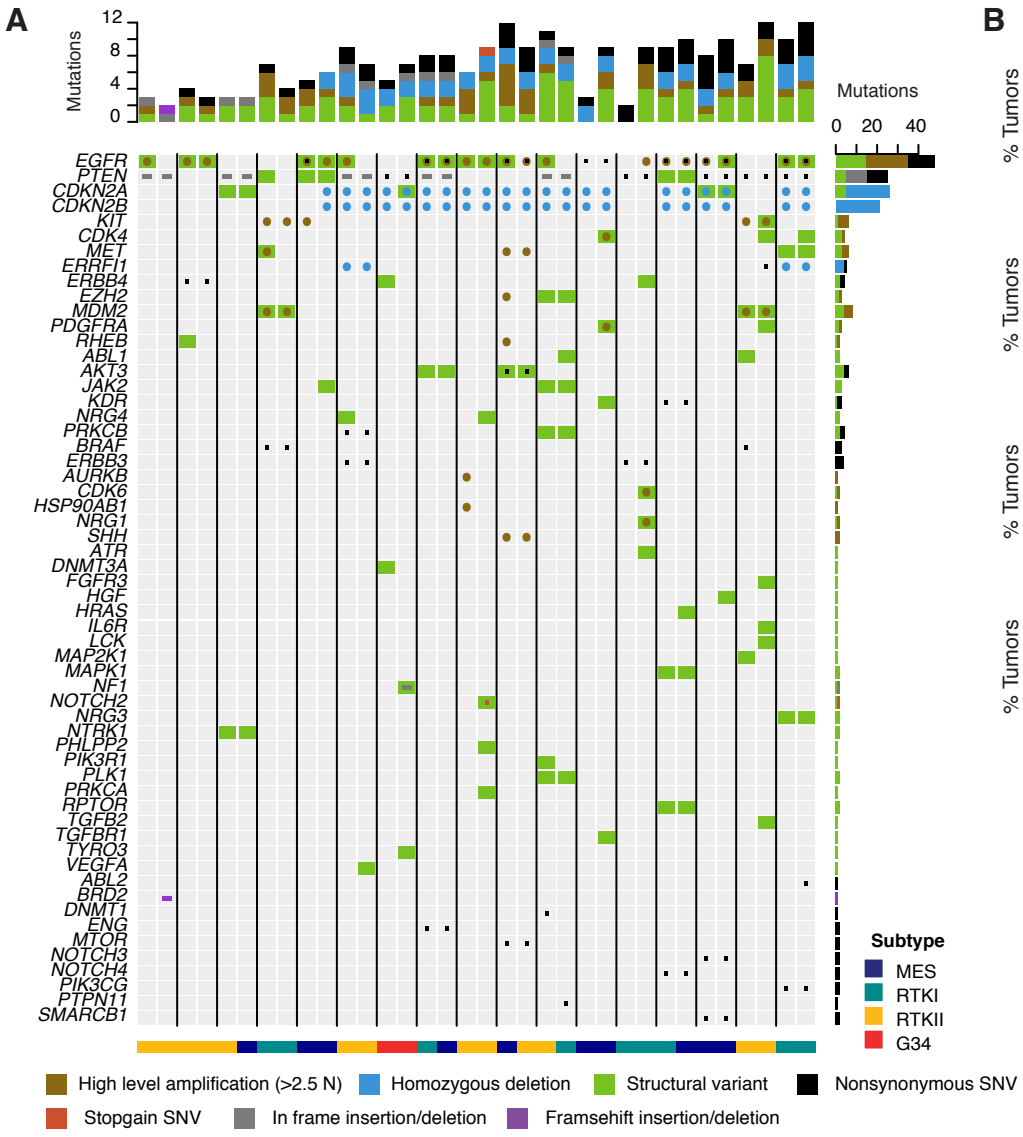


Figure S2. Mutational profile of primary and relapsed tumors, related to Figure 2.
(A) Potentially druggable mutations in primary and relapsed tumors of the discovery set (excluding hypermutated cases). Shown are mutations (non-synonymous/stop-gain SNVs, homozygous deletions, small insertions/deletions and structural variants) found in at least one tumor. Vertical lines separate tumor pairs (left primary, right relapsed); methylation subtypes are indicated at the bottom.
(B) Mutation distribution in the discovery and the validation set for four prominent oncogenes (*TERT* promoter, *PTEN*, *EGFR*, *TP53*; excluding samples analyzed with both methods).

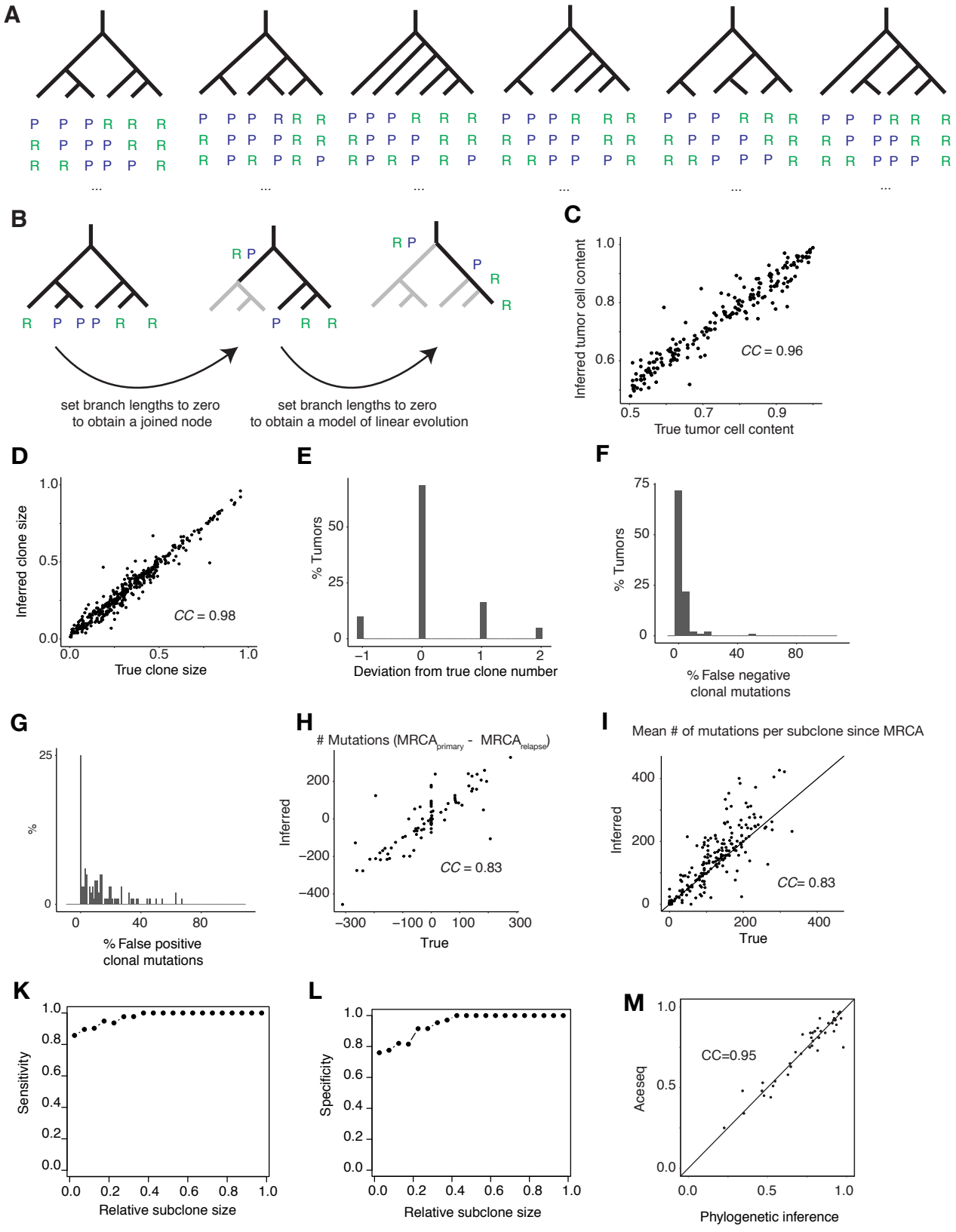

A
B
C
D
E
F
G
H
I
K
L
M
set branch lengths to zero
to obtain a joined node
set branch lengths to zero
to obtain a model of linear evolution
Inferred tumor cell content
True tumor cell content
CC = 0.96
Inferred clone size
True clone size
CC = 0.98
% Tumors
Deviation from true clone number
% False negative
clonal mutations
% False positive
clonal mutations
Mutations (MRCA primary - MRCA relapse)
Inferred
True
CC = 0.83
Mean # of mutations per subclone since MRCA
CC= 0.83
Sensitivity
Relative subclone size
Specificity
Relative subclone size
CC=0.95
Aceseq
Phylogenetic inference

Figure S3. Phylogenetic inference from whole genome sequencing data, related to Figure 3 and STAR Methods.

(A-B) Candidate trees for phylogenetic inference.

(A) Tree templates used to infer the phylogenetic structure. Three examples of unique distributions of primary and relapse samples are shown for each tree, if assuming maximally three subclones per sample.

(B) Trees can be collapsed to obtain solutions, in which primary and relapse samples share clones, or to topologies representing linear evolution.

(C-L) Phylogenetic inference on simulated data. Shown are the results from 100 simulations of up to three subclones per primary and relapse sample, respectively.

(C) True and inferred tumor cell content with Pearson's correlation coefficient.

(D) True and inferred clone size with Pearson's correlation coefficient. Shown are only cases, in which the correct number of subclones had been called.

(E) Deviation from the true number of subclones.

(F, G) False negative and false positive clonal mutations, if neglecting subclones ≤10%.

(H) True and inferred difference between the most recent common ancestors of primary and relapse sample, measured in mutation counts. Pearson's correlation coefficient is indicated in the plot.

(I) True and inferred mean number of mutations per subclone since the most recent common ancestor population of the tumor.

(K) Sensitivity of detecting a subclone in dependence of the relative subclone size.

(L) Specificity of detecting a subclone in dependence of the relative subclone size.

(M) Comparison of tumor cell content estimates from phylogenetic inference and from ACEseq. Pearson's correlation coefficient is indicated in the plot.

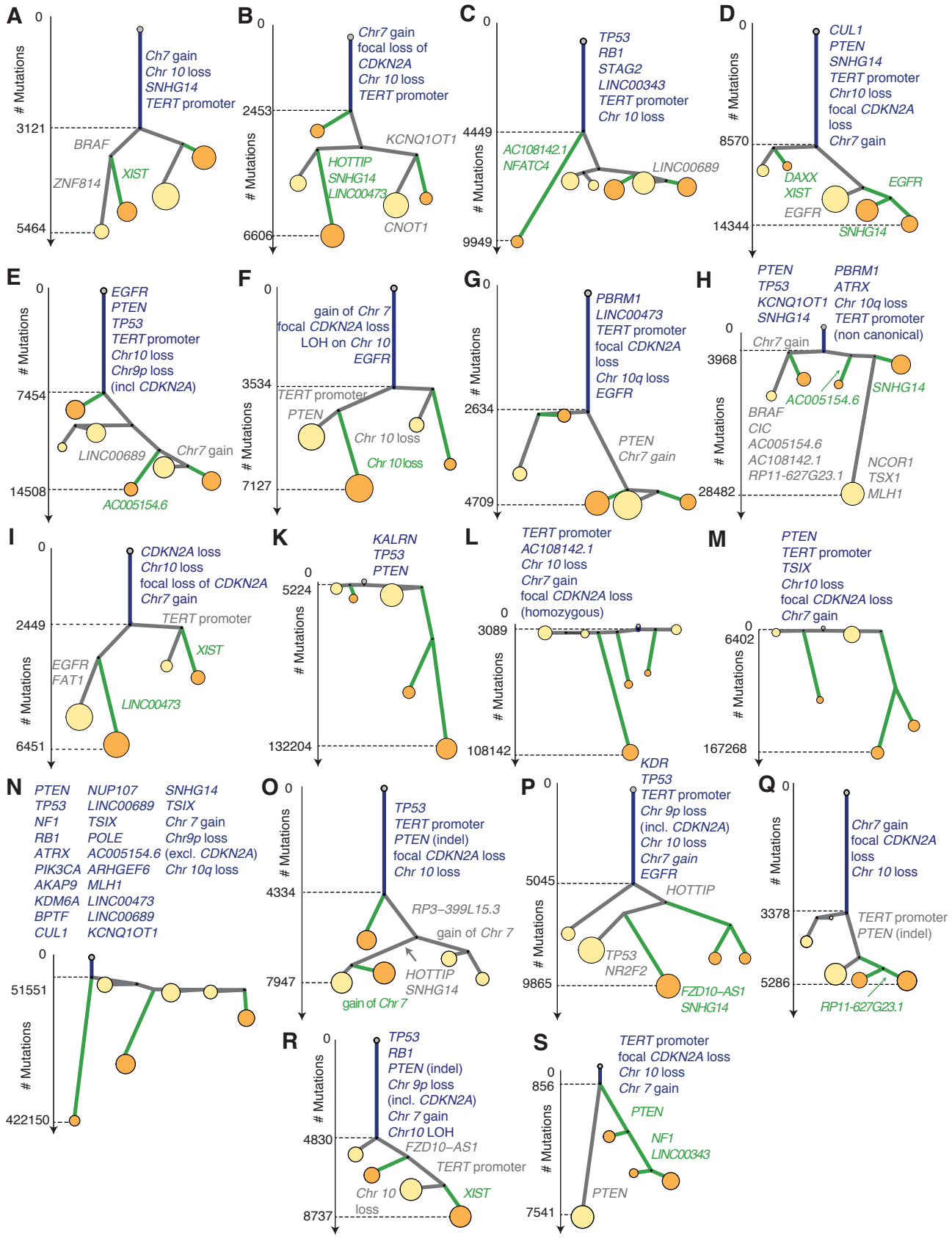

A
Mutations
0
3121
5464
Ch7 gain
Chr 10 loss
SNHG14
TERT promoter
BRAF
XIST
ZNF814
B
Mutations
0
2453
6606
Chr7 gain
focal loss of
CDKN2A
Chr 10 loss
TERT promoter
KCNQ1OT1
HOTTIP
SNHG14
LINC00473
CNOT1
C
Mutations
0
4449
9949
TP53
RB1
STAG2
LINC00343
TERT promoter
Chr 10 loss
AC108142.1
NFATC4
LINC00689
D
Mutations
0
8570
14344
CUL1
PTEN
SNHG14
TERT promoter
Chr10 loss
focal CDKN2A
loss
Chr7 gain
DAXX
XIST
EGFR
EGFR
SNHG14
E
Mutations
0
7454
14508
EGFR
PTEN
TP53
TERT promoter
Chr10 loss
Chr9p loss
(incl CDKN2A)
LINC00689
Chr7 gain
AC005154.6
F
Mutations
0
3534
7127
gain of Chr 7
focal CDKN2A loss
LOH on Chr 10
EGFR
TERT promoter
PTEN
Chr 10 loss
Chr 10 loss
G
Mutations
0
2634
4709
PBRM1
LINC00473
TERT promoter
focal CDKN2A
loss
Chr 10q loss
EGFR
PTEN
Chr7 gain
H
Mutations
0
3968
28482
PTEN
TP53
KCNQ1OT1
SNHG14
PBRM1
ATRX
Chr 10q loss
TERT promoter
(non canonical)
Chr7 gain
AC005154.6
SNHG14
BRAF
CIC
AC005154.6
AC108142.1
RP11-627G23.1
NCOR1
TSX1
MLH1
I
Mutations
0
2449
6451
CDKN2A loss
Chr10 loss
focal loss of CDKN2A
Chr7 gain
TERT promoter
XIST
EGFR
FAT1
LINC00473
K
Mutations
0
5224
132204
KALRN
TP53
PTEN
L
Mutations
0
3089
108142
TERT promoter
AC108142.1
Chr 10 loss
Chr7 gain
focal CDKN2A loss
(homozygous)
KDR
TP53
M
Mutations
0
6402
167268
PTEN
TERT promoter
TSIX
Chr10 loss
focal CDKN2A loss
Chr7 gain
N
PTEN
TP53
NF1
RB1
ATRX
PIK3CA
AKAP9
KDM6A
BPTF
CUL1
NUP107
LINC00689
TSIX
POLE
AC005154.6
ARHGEF6
MLH1
LINC00473
LINC00689
KCNQ1OT1
SNHG14
TSIX
Chr 7 gain
Chr9p loss
(excl. CDKN2A)
Chr 10q loss
Mutations
0
51551
422150
O
Mutations
0
4334
7947
TP53
TERT promoter
PTEN (indel)
focal CDKN2A loss
Chr 10 loss
RP3–399L15.3
gain of Chr 7
HOTTIP
SNHG14
gain of Chr 7
P
Mutations
0
5045
9865
TP53
TERT promoter
Chr 9p loss
(incl. CDKN2A)
Chr 10 loss
Chr7 gain
EGFR
HOTTIP
TP53
NR2F2
FZD10–AS1
SNHG14
Q
Mutations
0
3378
5286
Chr7 gain
focal CDKN2A
loss
Chr 10 loss
TERT promoter
PTEN (indel)
RP11-627G23.1
R
Mutations
0
4830
8737
TP53
RB1
PTEN (indel)
Chr 9p loss
(incl. CDKN2A)
Chr 7 gain
Chr10 LOH
FZD10–AS1
TERT promoter
Chr 10
loss
XIST
S
Mutations
0
856
7541
TERT promoter
focal CDKN2A loss
Chr 10 loss
Chr 7 gain
PTEN
NF1
LINC00343
PTEN

Figure S4. Phylogenetic trees of the discovery set, related to Figure 3.
Trees are designed as explained in Figure 3 (i.e. vertical branch lengths scale with the number of mutations, circle sizes scale with relative subclone sizes; mutations in driver genes and copy number changes of chromosomes 7, 9 and 10 are indicated; yellow circles, primary tumor; orange circles, recurrent tumor).
(A-N) Cases with an oligoclonal origin of the relapsed tumor. Panels (K-N) correspond to hypermutated cases, where only clonal mutations are indicated.
(O-S) Cases with a more monoclonal origin of the relapsed tumor. The tree shown in (Q) was due to low tumor cell content selected based on the purity estimate by ACEseq.
Shown are all tumors of the discovery set excluding the three tumors shown in Fig. 3.

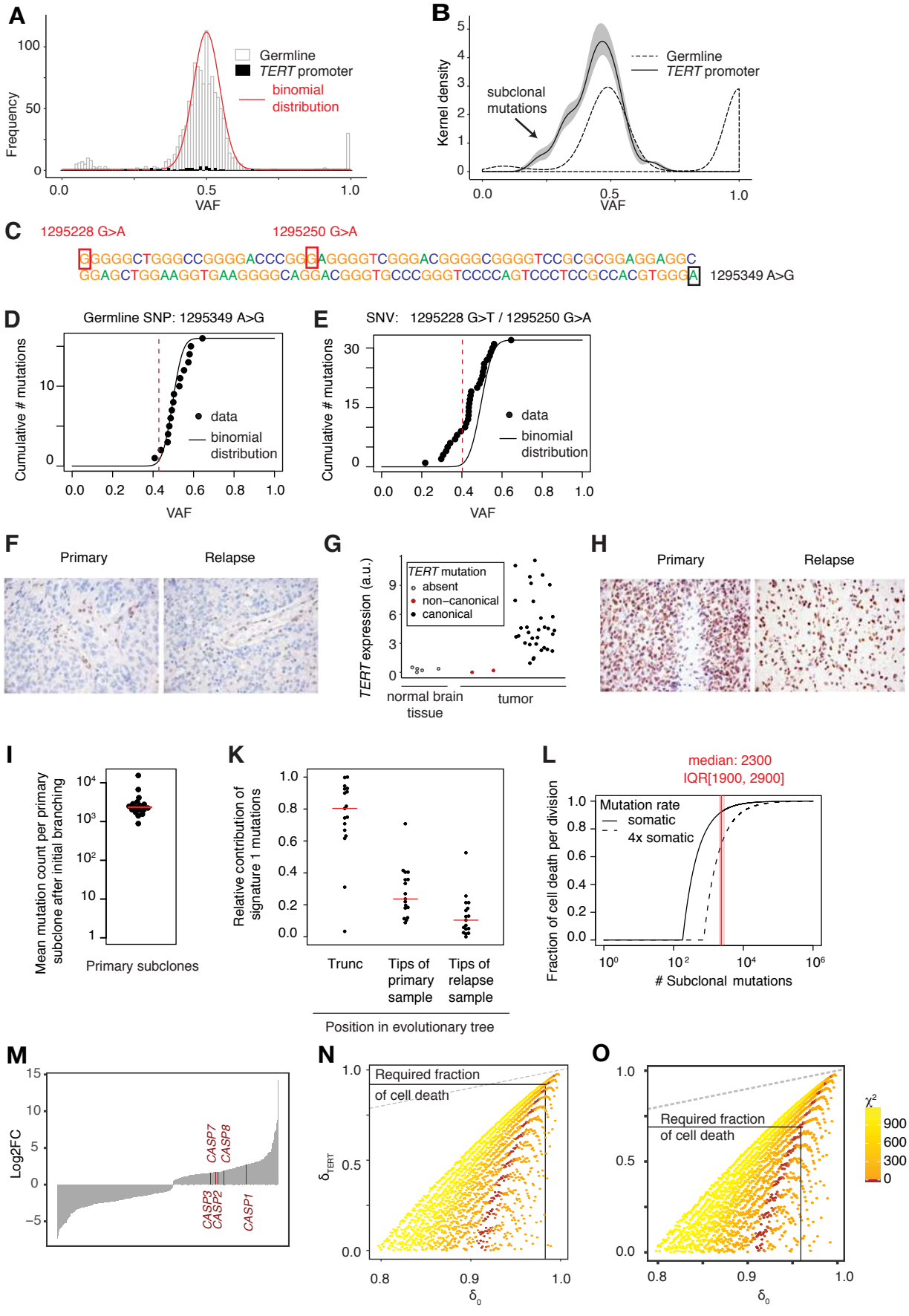

A
Germline
TERT promoter
binomial distribution
Frequency
VAF
B
Germline
TERT promoter
subclonal mutations
Kernel density
VAF
C
1295228 G>A
1295250 G>A
GGGGGCTGGGCCGGGGACCCGGGAGGGGTCGGGACGGGGCGGGGTCCGCGCGGAGGAGGC
GGAGCTGGAAGGTGAAGGGGCAGGACGGGTGCCCGGGTCCCCAGTCCCTCCGCCACGTGGGA
1295349 A>G
D
Germline SNP: 1295349 A>G
Cumulative # mutations
data
binomial distribution
VAF
E
SNV: 1295228 G>T / 1295250 G>A
Cumulative # mutations
data
binomial distribution
VAF
F
Primary
Relapse
G
TERT mutation
absent
non−canonical
canonical
TERT expression (a.u.)
normal brain tissue
tumor
H
Primary
Relapse
I
Mean mutation count per primary subclone after initial branching
Primary subclones
K
Relative contribution of signature 1 mutations
Trunc
Tips of primary sample
Tips of relapse sample
Position in evolutionary tree
L
median: 2300
IQR[1900, 2900]
Mutation rate
somatic
4x somatic
Fraction of cell death per division
Subclonal mutations
M
Log2FC
CASP7
CASP8
CASP3
CASP2
CASP1
N
Required fraction of cell death
δTERT
δ0
O
Required fraction of cell death
χ2
900
600
300
0
δ0

Figure S5. Subclonality of *TERT* promoter mutations and evolutionary dynamics during tumor growth, related to Figures 4 and 5.

(A) Variant allele frequencies (VAFs) of canonical *TERT* promoter mutations (black), germline mutations (grey) and expected VAFs if sampling from a binomial distribution (red). VAFs of *TERT* promoter mutations were adjusted with the average estimate of tumor cell content from phylogenetic inference and ACEseq (as shown in Fig. S3M; tumors with non-neutral copy number at the *TERT* promoter were excluded, yielding data from 16 tumor pairs). Germline mutations were chosen from all tumors of the discovery set at loci with read coverage within the inter-quartile-range of the coverage at the *TERT* promoter (IQR=[106,140], assessed from the 16 pairs of which the VAFs at the *TERT* promoter mutation are shown) and down-sampled to 1,000 loci for better visualization. Expected germline VAFs (red) were computed from the average of binomial distributions if drawing samples of the measured read coverages with success probability 0.5.

(B) Kernel density estimates of the VAFs at heterozygous germline mutations (grey) and *TERT* promoter mutations (black; the solid line shows the average, the opaque area the 95% confidence interval of the kernel density estimate after sampling the tumor cell contents 10,000 times from the estimates from phylogenetic inference and ACEseq; displayed data are downsampled as in (A)).

(C) Reference sequence at the *TERT* promoter (red, canonical SNVs; black, germline SNP).

(D,E) Measured (points) and predicted (lines) cumulative distributions of VAFs at canonical *TERT* promoter mutations and a germline SNP (solid lines, binomial distribution at the lower 25% quantile of the measured coverage; red lines, 95% lower confidence bound at the lower 25% quantile of the measured coverage).

(D) Heterozygous germline SNP (1295349 A>G) in the discovery set (median coverage=140x, IQR=[130x,156x]; displayed data are from eight tumor pairs in which the polymorphism was found).

(E) Canonical *TERT* promoter mutations in the discovery set (VAFs are corrected for tumor cell content; median coverage=97x, IQR=[77x,114x] after correcting for tumor cell content; tumors with non-neutral copy number at the *TERT* promoter were excluded, yielding data from 16 tumor pairs).

(F) Immunohistochemical staining for ATRX in a primary and relapsed tumor of a selected patient with a frame-shift *ATRX* mutation and a non-canonical *TERT* promoter mutation in both tumors (antibody binding visualized with diaminobenzidine as chromogen (brown), all sections counterstained with hematoxylin (blue), original microscopical magnification: 400x; in total samples from four patients were stained of which two are exemplarily shown in (F) and (H)). Note loss of nuclear ATRX expression in both tumor samples. Interspersed vascular endothelia serve as endogenous control with retained nuclear ATRX expression.

(G) *TERT* expression (RNA-seq read counts normalized by library size) in normal brain tissue samples (grey dots; normal brain tissue data were generated by The Cancer Genome Atlas (TCGA) research network: http://cancergenome.nih.gov/), a tumor pair with non-canonical *TERT* promoter mutation (red dots) and in tumors with canonical *TERT* promoter mutations (black dots). ATRX expression of the tumor pair with non-canonical *TERT* promoter mutation is shown in (F).

(H) Nuclear ATRX expression in a primary tumor without *ATRX* mutation and retention of nuclear ATRX expression in the relapsed tumor of the same patient with a non-synonymous SNV in *ATRX*. Staining and magnification as in (F).

(I) Mean number of mutations per primary subclone accumulated after the most recent common ancestor of the tumor (shown are the tumors of the discovery set, excluding hypermutated samples; red line, median).

(K) Relative contribution of the clock-like mutational signature AC1 to mutations residing at the trunk and the tips of the phylogenetic trees (separately for primary and relapse subclones from cases of the discovery set, excluding hypermutated samples; red lines, median).

(L) Fraction of cell death per cell division required to reconcile an observed number of subclonal mutations with a tumor size of 10^9 cells (red line, median number of mutations per subclone observed in the data if excluding hypermutated cases; shaded area, inter quartile range).

(M) Genes associated with positive regulation of cell death (GSEA: GO:0010942, from http://software.broadinstitute.org/gsea/msigdb/), which were differentially expressed ($p<0.01$ after Benjamini-Hochberg correction) between tumors (n=36) and normal brain tissue controls (n=5, normal brain tissue data were generated by The Cancer Genome Atlas (TCGA) research network: http://cancergenome.nih.gov/). Shown are log2 fold changes between tumor and control.

(N, O) Residual sum of squares when comparing the mean and variance of measured and modeled *TERT* promoter-mutant tumor fraction at different values of δ_0 and δ_{TERT} (grey dashed lines, bisetrices; data are from 19 tumors of the discovery set with canonical *TERT* promoter mutation). For each point the simulated mean and variance of 1000 simulations were compared to the measurement. Shown are the results if implementing

TERT promoter mutations at the somatic mutation rate (N) and four-fold the somatic mutation rate (O). Estimates lying within the 95% confidence interval (corresponding to a $\mathcal{X}^2 \leq 5.99$) are shaded in dark red. These parameter combinations give equally good fits. Black lines mark the fractions of cellular death required to explain tumor sizes of 10^9 cells (as shown in (L)). This fraction is required for both the *TERT* promoter-mutant tumor fraction and the tumor fraction without *TERT* promoter mutation. Since $\delta_{TERT} \leq \delta_0$, it is sufficient to require the cutoff for δ_{TERT}, corresponding to the horizontal line. Thus parameter estimates explaining both, the *TERT* promoter-mutant tumor fraction and mutation accumulation in tumors of realistic sizes, intersect the horizontal black line and are colored in dark red. The corresponding values for δ_0 are marked by vertical lines.